Challenges of Trustable AI and Added-Value on Health

B. Séroussi, Patrick Weber, Ferdinand Dhombres, Cyril Grouin, Jan-David Liebe, Sylvia Pelayo, Andrea Pinna, Bastien Rance, Lucia Sacchi, Adrien Ugon, et al.

STUDIES IN HEALTH TECHNOLOGY AND INFORMATICS 294

Challenges of Trustable AI and Added-Value on Health

Proceedings of MIE 2022

Editors:
Brigitte Séroussi
Patrick Weber
Ferdinand Dhombres
Cyril Grouin
Jan-David Liebe
Sylvia Pelayo
Andrea Pinna
Bastien Rance
Lucia Sacchi
Adrien Ugon
Arriel Benis
Parisis Gallos

IOS Press

Artificial Intelligence (AI) in healthcare promises to improve the accuracy of diagnosis and screening, support clinical care, and assist in various public health interventions such as disease surveillance, outbreak response, and health system management. But the increasing importance of AI in healthcare means that trustworthy AI is vital to achieve the beneficial impacts on health anticipated by both health professionals and patients.

This book presents the proceedings of the 32nd Medical Informatics Europe Conference (MIE2022), organized by the European Federation for Medical Informatics (EFMI) and held from 27 - 30 May 2022 in Nice, France. The theme of the conference was *Challenges of Trustable AI and Added-Value on Health*. Over 400 submissions were received from 43 countries, and were reviewed in a thorough process by at least three reviewers before being assessed by an SPC co-chair, with papers requiring major revision undergoing further review. Included here are 147 full papers (acceptance rate 54%), 23 short papers and 79 posters from the conference. Topics covered include the usual sub-domains of biomedical informatics: decision support and clinical information systems; clinical research informatics; knowledge management and representation; consumer health informatics; natural language processing; public health informatics; and privacy, ethical and societal aspects, but also innovative approaches to the collection, such as organization and analysis of data and knowledge related to health and wellbeing, as well as theoretical and applied contributions to AI methods and algorithms.

Providing an overview of the latest developments in medical informatics, the book will be of interest to all those involved in the development and provision of healthcare today.

ISBN 978-1-64368-284-6 (print)
ISBN 978-1-64368-285-3 (online)
ISSN 0926-9630 (print)
ISSN 1879-8365 (online)

CHALLENGES OF TRUSTABLE AI AND ADDED-VALUE ON HEALTH

Studies in Health Technology and Informatics

International health informatics is driven by developments in biomedical technologies and medical informatics research that are advancing in parallel and form one integrated world of information and communication media and result in massive amounts of health data. These components include genomics and precision medicine, machine learning, translational informatics, intelligent systems for clinicians and patients, mobile health applications, data-driven telecommunication and rehabilitative technology, sensors, intelligent home technology, EHR and patient-controlled data, and Internet of Things.

Studies in Health Technology and Informatics (HTI) series was started in 1990 in collaboration with EU programmes that preceded the Horizon 2020 to promote biomedical and health informatics research. It has developed into a highly visible global platform for the dissemination of original research in this field, containing more than 250 volumes of high-quality works from all over the world.

The international Editorial Board selects publications with relevance and quality for the field. All contributions to the volumes in the series are peer reviewed.

Volumes in the HTI series are submitted for indexing by MEDLINE/PubMed; Web of Science: Conference Proceedings Citation Index – Science (CPCI-S) and Book Citation Index – Science (BKCI-S); Google Scholar; Scopus; EMCare.

Volume 294

Recently published in this series

Vol. 293 G. Schreier, B. Pfeifer, M. Baumgartner and D. Hayn (Eds.), dHealth 2022 – Proceedings of the 16th Health Informatics Meets Digital Health Conference
Vol. 292 T. Bürkle, K. Denecke, J. Holm, M. Sariyar and M. Lehmann (Eds.), Healthcare of the Future 2022 – Digital Health – From Vision to Best Practice!
Vol. 291 T.M. Deserno, M. Haghi and N. Al-Shorbaji (Eds.), Accident and Emergency Informatics
Vol. 290 P. Otero, P. Scott, S.Z. Martin and E. Huesing (Eds.), MEDINFO 2021: One World, One Health – Global Partnership for Digital Innovation – Proceedings of the 18th World Congress on Medical and Health Informatics
Vol. 289 J. Mantas, A. Hasman, M.S. Househ, P. Gallos, E. Zoulias and J. Liaskos (Eds.), Informatics and Technology in Clinical Care and Public Health

ISSN 0926-9630 (print)
ISSN 1879-8365 (online)

Challenges of Trustable AI and Added-Value on Health

Proceedings of MIE 2022

Edited by

Brigitte Séroussi
Sorbonne University, INSERM, Laboratory of Medical Informatics and Knowledge Engineering in e-Health (LIMICS); AP-HP, Paris, France
Patrick Weber
Nice Computing, Lausanne, Switzerland
Ferdinand Dhombres
Sorbonne University, INSERM, Laboratory of Medical Informatics and Knowledge Engineering in e-Health (LIMICS); AP-HP, Paris, France
Cyril Grouin
University Paris-Saclay, CNRS, Interdisciplinary Laboratory of Digital Sciences (LISN), Orsay, France
Jan-David Liebe
Medical School Hamburg, Department for Digital Health Management, Hamburg, Germany; University of Applied Sciences, Health Informatics Research Group, Osnabrück, Germany; UMIT, Institute for Medical Informatics, Hall in Tirol, Austria
Sylvia Pelayo
Clinical Investigation Center for Innovative Technologies (CIC-IT), Lille University, Lille, France
Andrea Pinna
Sorbonne University, CNRS, Paris, France
Bastien Rance
University Paris Cité, INSERM, Paris, France; HeKA, Inria, Paris, France
Lucia Sacchi
Laboratory for Biomedical Informatics, Department of Electrical, Computer, and Biomedical Engineering, University of Pavia, Italy
Adrien Ugon
Sorbonne University, CNRS, Paris, France; Univ Gustave Eiffel, ESIEE Paris, Département SEN, Noisy-le-Grand, France
Arriel Benis
Faculty of Industrial Engineering and Technology Management, and Faculty of Digital Technologies in Medicine, Holon Institute of Technology, Israel
Parisis Gallos
Computational Biomedicine Research Lab, Department of Digital Systems, University of Piraeus, Greece

IOS Press

Amsterdam • Berlin • Washington, DC

ISBN 978-1-64368-284-6 (print)
ISBN 978-1-64368-285-3 (online)
Library of Congress Control Number: 2022938531
doi: 10.3233/SHTI294

Publisher
IOS Press BV
Nieuwe Hemweg 6B
1013 BG Amsterdam
Netherlands
fax: +31 20 687 0019
e-mail: order@iospress.nl

For book sales in the USA and Canada:
IOS Press, Inc.
6751 Tepper Drive
Clifton, VA 20124
USA
Tel.: +1 703 830 6300
Fax: +1 703 830 2300
sales@iospress.com

Preface

The 32nd Medical Informatics Europe Conference, MIE2022, was held in Nice, France, from 27 to 30 May 2022. The Conference was hosted by the European Federation for Medical Informatics (EFMI) and organized by the "MCO Congress". The Scientific Programme Committee was chaired by Professor Brigitte Séroussi, Sorbonne University, French Association of Medical Informatics (AIM). The overarching theme of MIE2022 was "Challenges of Trustable AI and Added-value on Health" stressing the increasing importance of Artificial Intelligence (AI) in healthcare on the one hand and the need for a trustworthy AI on the other hand, in order to reach the full expected impact of AI on health from both health-professional and patient-centered perspectives.

Developed in the 1970s, the first AI systems were essentially knowledge-based decision support systems. Despite their good performance, these first knowledge-based systems were never routinely used on real patients, and because they were able to explain their reasoning process, most of them turned out to be more beneficial for teaching than for clinical practice. After some "winters", AI is back, with new machine learning methods that promise to improve the accuracy of diagnosis and screening, support clinical care, and assist various public health interventions such as disease surveillance, outbreak response, and health system management. Naturally, as these new AI systems emerge, concerns arise concerning the level of control that should be conceded when reviewing the pace at which AI methods are introduced. One concern in particular arises from the fact that these new AI systems often work as "black boxes", and are unable to explain their results. Explainability is critical, however, to respond to patient and practitioner narrative exchanges, and the fact that practitioners, who are responsible for their decisions, cannot easily follow the proposals of AI systems that they disagree with is also a problem.

Throughout this publication, readers will find innovative approaches to the collection, organization, and analysis of data and knowledge related to health and wellbeing, as well as theoretical and applied contributions to AI methods and algorithms. Papers covering the usual subdomains of biomedical informatics (decision support systems, clinical information systems, clinical research informatics, knowledge management and representation, consumer health informatics, natural language processing, bioinformatics, public health informatics, privacy, ethical and societal aspects, etc.) are also offered. The Proceedings are published as an e-book by IOS Press in the series Studies in Health Technology and Informatics (HTI), providing open access for ease of use and browsing without the loss of any of the advantages of indexing and citation, in the major Scientific Literature Databases, such as Medline and Scopus.

Paris, 20.04.2022

The Editors,
Brigitte Séroussi, Patrick Weber, Ferdinand Dhombres, Cyril Grouin, Jan-David Liebe, Sylvia Pelayo, Andrea Pinna, Bastien Rance, Lucia Sacchi, Adrien Ugon, Arriel Benis, Parisis Gallos

This page intentionally left blank

About the Conference

The Conference

The 32nd Medical Informatics Europe Conference (MIE2022), organized by the European Federation for Medical Informatics Association (EFMI), was held from 27–30 May 2022 in Nice, France. Conference management was by "MCO Congress" (https://mcocongres.com/en/), endorsed by the French Association of Medical Informatics (AIM, https://www.france-aim.org/page/79803-l-aim).

Founded in 1976, EFMI (https://efmi.org/) is the leading organization in medical informatics in Europe, representing 30 countries through their respective national health informatics associations. EFMI is a not-for-profit organization concerned with the theory and practice of information science and technology within healthcare, and health sciences in a European context.

MIE is a medical informatics conference aimed at promoting research and development in biomedical and health informatics. The conference proposes scientific sessions consisting of oral presentations of peer-reviewed, full papers and short communication papers. The conference also includes panels, workshops, demos, and tutorials, some of which have been prepared by EFMI working groups. A large exhibition of peer-reviewed posters also forms part of the conference.

In 2022, the special theme of the conference was "Challenges of Trustable AI and Added-value on Health" (https://mie2022.org/).

Relevant topics included:

- Special Topic: Challenges of trustable AI and added-value on health
- Bioinformatics
- Citizen health informatics
- Clinical information systems
- Decision-support systems
- Education
- Health information systems
- Human factors and organizational issues
- Knowledge and information representation and modeling
- Medical robotics
- Natural language processing
- Patient records
- Public health and epidemiology informatics
- Security and safety
- Sensors, signals and imaging informatics
- Societal aspects
- Telehealth
- Visualization

MIE2022 conference included four keynotes by internationally recognized experts in medical informatics:

- **"Prediction Models for Personalized Preventative Medicine: Past, Present, Future"**, by Niels Peek, Professor of Health Informatics at the University of Manchester and director of the Christabel Pankhurst Institute for Health Technology Research and Innovation. Niels Peek also co-leads the International Centre for Translational Digital Health, and is a fellow of the Alan Turing Institute.
- **"The Evolution of AI in Medicine: How the Past Informs the Future"**, by Edward H (Ted) Shortliffe, Chair Emeritus and Adjunct Professor of Biomedical Informatics at Columbia University. Editor Emeritus of JBI. Textbook: Biomedical Informatics: Computer Applications in Health Care and Biomedicine (5th edition, 2021). ACM Grace Murray Hopper Award (1976), ACMI Morris F. Collen Award (2006), IMIA François Grémy Award of Excellence (2021).
- **"Patient-Centered Artificial Intelligence: Challenges and Opportunities"**, by Carolyn Petersen, patient and consumer advocate, Senior Editor of mayoclinic.org, the Mayo Clinic's health information website (USA).
- **"Building Trustworthy AI Systems with Reliable Components"**, by Riccardo Bellazzi, Professor of Bioengineering and Biomedical Informatics at the University of Pavia (Italy), IAHSI and ACMI Fellow, expert in AI in bio-medicine with a focus on data and knowledge integration (editorial board member of biomedical informatics journals and active in IMIA, AMIA, EFMI and in the Italian Society of Biomedical Informatics (SIBIM)).

Scientific Program Committee / Chair and co-chairs (alphabetic)

- Prof. Brigitte Séroussi (chair), Sorbonne University, INSERM, Laboratory of Medical Informatics and Knowledge Engineering in e-Health (LIMICS); AP-HP, Paris, France.
- Dr. Ferdinand Dhombres (co-chair), Sorbonne University, INSERM, Laboratory of Medical Informatics and Knowledge Engineering in e-Health (LIMICS); GRC26, AP-HP, Paris, France.
- Dr. Cyrille Grouin (co-chair), University Paris-Saclay, CNRS, Interdisciplinary Laboratory of Digital Sciences (LISN), Orsay, France.
- Prof. Jan-David Liebe (co-chair), Medical School Hamburg, Department for Digital Health Management, Hamburg, Germany; University of Applied Sciences, Health Informatics Research Group, Osnabrück, Germany; UMIT, Institute for Medical Informatics, Hall in Tirol, Austria.
- Dr. Sylvia Pelayo (co-chair), Clinical Investigation Center for Innovative Technologies (CIC-IT) 1403, Lille University, Lille, France.
- Dr. Andrea Pinna (co-chair), Sorbonne University, CNRS, LIP6, Paris, France
- Dr. Bastien Rance (co-chair), University Paris Cité, INSERM, UMRS 1138, F-75006 Paris, France; HeKA, Inria, F-75015 Paris, France.

- Dr. Lucia Sacchi (co-chair), Laboratory for Biomedical Informatics, Department of Electrical, Computer, and Biomedical Engineering, University of Pavia, Italy.
- Dr. Adrien Ugon (co-chair), Sorbonne University, CNRS, LIP6, Paris, France; Univ Gustave Eiffel, ESIEE Paris, Département SEN, F-93162, Noisy-le-Grand, France.
- Dr. Patrick Weber (co-chair), Nice Computing, Lauzanne, Switzerland.

Editorial Committee / Co-chairs (alphabetic)

- Dr. Arriel Bennis (co-chair), Faculty of Industrial Engineering and Technology Management, and Faculty of Digital Technologies in Medicine, Holon Institute of Technology, Israel.
- Dr. Parisis Gallos (co-chair), Computational Biomedicine Research Lab, Department of Digital Systems, University of Piraeus, Greece.

Peer Review Process

We received over 400 submissions from 43 countries. A thorough review process was conducted with valuable support from the 300 active reviewers. Almost all submissions were reviewed by at least three reviewers and assessed by one SPC co-chair. Final decisions were made based on their recommendations by SPC members during a two-day, face-to-face meeting in Paris, France. Papers requiring major revisions underwent a further peer review.

At the end of this process 39 submissions were retracted, and from the 271 full papers, 37 short communication papers, 33 posters, nine demonstrations, 13 panels, 17 workshops and five tutorials submitted, 147 papers (acceptance rate of 54%), 23 short communication papers (conversion of 11 full papers), 79 posters (conversion of 51 full papers and 13 short communication papers), eight demonstrations, nine panels, 16 workshops, and five tutorials were accepted. All accepted full papers, short communication papers, and posters are included in these proceedings.

We would like to take this opportunity to thank the SPC co-chairs and all the reviewers for their invaluable contribution to MIE2022.

Sponsors (as of April 25th, 2022)

Silver Sponsor

- Norwegian Centre for E-health Research, Tromso, Norway

Partners

- Health Data Hub, Paris, France
- HL7 Europe, Brussels, Belgium

Address for correspondence: Brigitte.seroussi@aphp.fr

This page intentionally left blank

Contents

Section IV. Decision Support Systems and Patient Records

Section I

Challenges of Trustable AI and Added-Value on Health

This page intentionally left blank

Challenges of Trustable AI and Added-Value on Health
B. Séroussi et al. (Eds.)

doi:10.3233/SHTI220385

Applying Machine Learning to Arsenic Species and Metallomics Profiles of Toenails to Evaluate Associations of Environmental Arsenic with Incident Cancer Cases

Sheida MAJOUNI[a], Jong Sung KIM[b,c], Ellen SWEENEY[d], Erin KELTIE[b,c] and Syed Sibte Raza ABIDI[a]
[a] *NICHE Research Group, Faculty of Computer Science, Dalhousie University, Halifax, Canada*
[b] *Department of Community Health and Epidemiology, Faculty of Medicine, Dalhousie University, Halifax, Nova Scotia, Canada*
[c] *Health and Environments Research Centre (HERC) Laboratory, Faculty of Medicine, Dalhousie University, Halifax, Nova Scotia, Canada*
[d] *Atlantic Partnership for Tomorrow's Health (PATH), Faculty of Medicine, Dalhousie University, Halifax, Nova Scotia, Canada*

Abstract. Chronic exposure to environmental arsenic has been linked to a number of human diseases affecting multiple organ systems, including cancer. The greatest concern for chronic exposure to arsenic is contaminated groundwater used for drinking as it is the main contributor to the amount of arsenic present in the body. An estimated 40% of households in Nova Scotia (Canada) use water from private wells, and there is a concern that exposure to arsenic may be linked to/associated with cancer. In this preliminary study, we are aiming to gain insights into the association of environmental metal's pathogenicity and carcinogenicity with prostate cancer. We use toenails as a novel biomarker for capturing long-term exposure to arsenic, and have performed toxicological analysis to generate data about differential profiles of arsenic species and the metallome (entirety of metals) for both healthy and individuals with a history cancer. We have applied feature selection and machine learning algorithms to arsenic species and metallomics profiles of toenails to investigate the complex association between environmental arsenic (as a carcinogen) and prostate cancer. We present machine learning based models to ultimately predict the association of environmental arsenic exposure in cancer cases.

Keywords. Cancer Prediction, Machine Learning, Arsenic Exposure.

1. Introduction

Chronic exposure to environmental arsenic has been observed to affect multiple human organs leading to chronic diseases such as diabetes, cancer [1], cardiovascular and neurological disorders. Chronic arsenic exposure occurs primarily through contaminated groundwater used for drinking, especially water drawn from wells in areas with naturally

occurring environmental arsenic. In Atlantic Canada, the province of Nova Scotia (NS) has high levels of naturally occurring environmental arsenic [2]. An estimated 40% of households in NS use well water, and 12% of wells were found to have arsenic concentrations above the Canadian drinking water safety guideline. Furthermore, Atlantic Canada has the highest incidence rates of all cancers in the country. Given Atlantic Canada's high cancer rates, presence of environmental arsenic and use of well water, it is important to understand how exposure to environmental arsenic contributes to the incidence of cancer in the region.

This long-term research study is investigating the impact of environmental arsenic's pathogenicity and carcinogenicity—i.e., the arsenic body burden which measures the amount of arsenic present in the body—on the incidence of different types of cancers. We aim to understand the associations between environmental arsenic exposure and cancer types by using multiple biomarkers—i.e., urine that captures short-term arsenic exposure and toenails that capture long-term chronic arsenic exposure. The distinguishing aspect of our approach is the use of toenails as a new biomarker for environmental metal exposure, and the use of arsenic speciation in addition to metallomics profiles [3]. The biomarkers for both cancer and healthy participants were collected by Atlantic Partnership for Tomorrow's Health (Atlantic PATH) which has the world's largest biobank of toenails [4]. We performed toxicological analysis of these biomarkers, collected from both individuals who are healthy and with a history of cancer, to generate complete arsenic species and metallomics profiles. Using the arsenic species and metallomics profiles, we applied Machine Learning (ML) methods to generate cancer prediction models to differentiate between healthy individuals and those with a history of prostate cancer based on toxicological analysis of their toenail samples.

2. Background

Most studies assessing speciation have focused on analysis of urine which is better suited for assessing short-term exposure (3-4 days). Since many arsenic-related diseases arise from chronic exposure, measuring longer-term biomarkers may prove to be a more relevant. Since toenails grow slowly they can provide a longer period (2-12 months) of arsenic exposure data [3] as such are more suited as a novel biomarker to study the molecular mechanisms of metal pathogenicity and carcinogenicity.

ML methods have been used to predict disease risk using the metallome profile. Ahmed and Santosh [5] used the concentrations of metal elements from a serum metallomic analysis to predict the likelihood of Parkinson's disease using with a neural network. Luo [6] used Bayesian kernel machine regression to examine associations between heavy metals (blood levels) and indicators of kidney functions, where exposure of a co-occurring heavy metals mixture was associated with poor kidney function. Tan et al. [7] applied Adaboost to trace metal elements from urine samples to predict early lung cancer. Guo et al. [8] applied support vector machine to metal elements from hair samples to predict prostate cancer. These studies in the literature show the potential of using ML algorithms to predict disease risk based on metallome profiles, it is noted that these models were developed using a small dataset (maximum 160 samples), conventional biomarkers and simple ML algorithms.

3. Methodology

The data was provided by Atlantic PATH, a longitudinal cohort study in Atlantic Canada with more than 35,000 participants. Participants provided questionnaire data, physical measures and biosamples [4]. The dataset contained 573 individuals—431 healthy controls and 142 participants with a history of cancer. The data include demographics, arsenic species and metallome profiles of all metals. This data and the toenails were collected between 2009-2015 during baseline recruitment.

3.1. Toxicological Analysis

Toxicological analysis of the biomarkers was performed by the Health and Environments Research Centre (HERC) Laboratory at Dalhousie University. High Performance Liquid Chromatography (HPLC) and Inductively Coupled Plasma-Mass Spectrometry (ICP-MS) was used for analysis of toenails samples. For arsenic speciation, three forms were measured: (1) inorganic arsenic (iAs: arsenite + arsenate), (2) monomethylarsonic acid (MMAV), and (3) dimethylarsinic acid (DMA). Separately, the concentrations of 36 additional metals (e.g., Cd, Ni, Pb, etc.) as a metallomics approach were measured. The toxicological analysis of toenails yielded a dataset of arsenic species as %MMA, %DMA, %iAs, PMI, SMI and metallome profiles of all metals.

3.2. Developing Classification Models using Machine Learning Methods

This preliminary study is based on a small dataset, large number of attributes and a class imbalance. Given these challenges, our methodology included steps to select important features, address the class imbalance and use different types of ML methods to predict association between environmental arsenic and cancer. We experimented with the following ML methods: Decision tree (DT), random forest (RF), Support Vector Machine (SVM) with the various kernels, Neural Network (NN), ensemble methods like XG-boost, and deep learning methods.

4. Results

Baseline classification experiments using the original data yielded a low classification accuracy in the 70-75% range. Next, to improve classification accuracy we addressed the class imbalance issue as the cancer class is underrepresented. We applied Adaptive Synthetic Sampling Method (ADASYN) and SMOTE, however we the up-sampling experiment did not improve the accuracy as shown in Table 1.

Table 1. Classification accuracy of ML models using up-sampling methods.

ML Methods	SMOTE	ADASYN
DT	71.23% ($F_{1\text{-score}}$: 0.23)	64.8% ($F_{1\text{-score}}$: 0.31)
RF	**79.79%($F_{1\text{-score}}$: 0.47)**	**72.5% ($F_{1\text{-score}}$: 0.60)**
5-NN	56.00% ($F_{1\text{-score}}$: 0.46)	59.48% ($F_{1\text{-score}}$: 0.50)
GNB	38.27% ($F_{1\text{-score}}$: 0.37)	42.10% ($F_{1\text{-score}}$: 0.36)
XG-Boost	72.13% ($F_{1\text{-score}}$: 0.19)	62.00% ($F_{1\text{-score}}$: 0.45)

To investigate feature importance, we applied dimension reduction methods of PCA, LDA, and parallel axis plot (Figure 1) which indicated that Zn, Se, pDMA and SMI can be removed. Table 2 shows the classification results. The classification accuracy for the RF increased from 72.5% to 84.99% by parallel axis dimension reduction method.

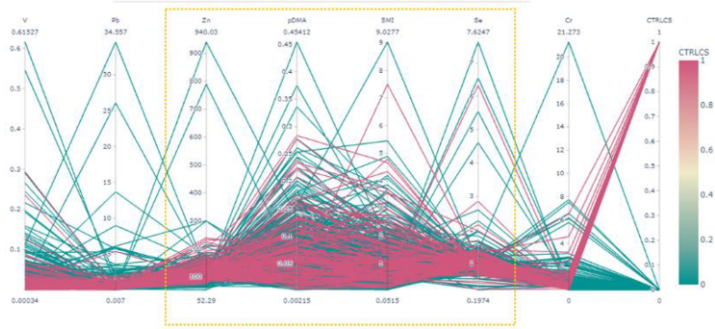


Figure 1. Red lines and blue ones indicate cancer samples and healthy people respectively. Note that samples of two classes are distributed all over Zn, pDMA, SMI, and Se axis.

Table 2. Classification accuracy results after dimension reduction

ML Methods	LDA	PCA	Parallel axis
DT	70.51%	71.73%	68.2% ($F_{1\text{-score}}$: 0.18)
RF	**76.96%**	**78.36%**	**84.9%($F_{1\text{-score}}$: 0.86)**
5-NN	68.9%	69.98%	67% ($F_{1\text{-score}}$: 0.14)
GNB	74.7%	65.79%	38% ($F_{1\text{-score}}$: 0.37)
All SVM Methods	75%	75%	75% ($F_{1\text{-score}}$: 0.0)
XG-Boost	75%	75%	75% ($F_{1\text{-score}}$: 0.0)

Table 3. Classification accuracy results using the autoencoder

Methods	Accuracy
DT	90.05% ($F_{1\text{-score}}$: 0.80)
RF	**97.56% ($F_{1\text{-score}}$: 0.96)**
5-NN	91.8% ($F_{1\text{-score}}$: 0.80)
GNB	92.15% ($F_{1\text{-score}}$: 0.81)
XG-Boost	84.64% ($F_{1\text{-score}}$: 0.56)
RBF SVM	94.06% ($F_{1\text{-score}}$: 0.87)

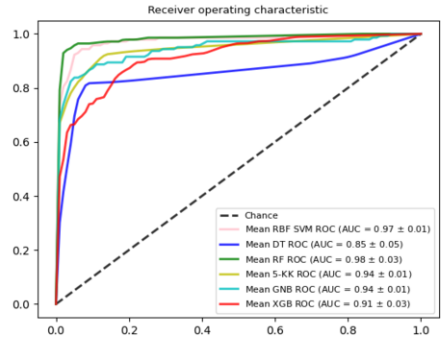


Figure 2. ROC for the final classification model

In the last experiment, we applied an autoencoder to find nonlinear relations between features for dimensional reduction. We used a deep autoencoder consisting of 4 hidden layers—the number of neurons in the encoder section was equal to 18, 9, 4, and 2—the third layer was used for training the final set of prediction models. Table 3 presents the

classification accuracy and when compared with Table 2, the classification accuracy for all the ML models significantly improved with the of the use of the autoencoder.

5. Discussion

The main goal of this broader study is to investigate the association between environmental exposure of carcinogens and cancer risk. The use of ML methods to analyze toxicological data derived from novel biomarkers offers a new approach to investigate the risks for multiple cancer types (including prostate, breast, skin cancers) associated with environmental exposure of arsenic and other metals. This preliminary analysis demonstrated the potential of ML based data analysis for predicting the association between arsenic exposure and prostate cancer. Future research will build upon this work whereby we plan to analyze broader population cohorts resulting in large datasets to develop ML models to predict the likelihood of different cancer types associated with toxic metals exposure to formulate population level cancer risk prevention strategies.

Acknowledgements

This research was conducted using Atlantic PATH data and biosamples with funding from the NS Health Research Fund and the Canadian Cancer Society/New Brunswick Health Research Foundation. The data used in this research were made available by the Atlantic PATH study, which is the Atlantic Canada regional component of the CanPath funded by the Canadian Partnership Against Cancer and Health Canada. The views expressed herein represent the views of the authors and do not necessarily represent the views of Health Canada.

References

[1] Martinez VD, Vucic EA, Becker-Santos DD, Gil L, Lam WL. Arsenic Exposure and the Induction of Human Cancers. Journal of Toxicology; 2011:1–13. https://doi.org/10.1155/2011/431287.

[2] Chappells H, Campbell N, Drage J, Fernandez CV, Parker L, Dummer TJB. Understanding the translation of scientific knowledge about arsenic risk exposure among private well water users in Nova Scotia. Science of The Total Environment 2015; 505:1259–73. https://doi.org/10.1016/j.scitotenv.2013.12.108.

[3] Signes-Pastor AJ, Gutiérrez-González E, García-Villarino M, Rodríguez-Cabrera FD, López-Moreno JJ, Varea-Jiménez E, et al. Toenails as a biomarker of exposure to arsenic: A review. Environmental Research 2021; 195:110286. https://doi.org/10.1016/j.envres.2020.110286.

[4] Sweeney E, Cui Y, DeClercq V, Devichand P, Forbes C, Grandy S, et al. Cohort Profile: The Atlantic Partnership for Tomorrow's Health (Atlantic PATH) Study. International Journal of Epidemiology 2017; 46:1762–1763i. https://doi.org/10.1093/ije/dyx124.

[5] Ahmed SSSJ, Santosh W. Metallomic Profiling and Linkage Map Analysis of Early Parkinson's Disease: A New Insight to Aluminum Marker for the Possible Diagnosis. PLoS ONE 2010; 5:e11252. https://doi.org/10.1371/journal.pone.0011252.

[6] Luo J, Hendryx M. Metal mixtures and kidney function: An application of machine learning to NHANES data. Environmental Research 2020; 191:110126. https://doi.org/10.1016/j.envres.2020.110126.

[7] Tan C, Chen H, Xia C. Early prediction of lung cancer based on the combination of trace element analysis in urine and an Adaboost algorithm. Journal of Pharmaceutical and Biomedical Analysis 2009; 49:746–52. https://doi.org/10.1016/j.jpba.2008.12.010.

[8] Guo J, Deng W, Zhang L, Li C, Wu P, Mao P. Prediction of prostate cancer using hair trace element concentration and support vector machine method. Biol Trace Elem Res 2007; 116:257–71.

Challenges of Trustable AI and Added-Value on Health
B. Séroussi et al. (Eds.)

doi:10.3233/SHTI220386

User Satisfaction with an AI System for Chest X-Ray Analysis Implemented in a Hospital's Emergency Setting

Diego RABINOVICH[a], Candelaria MOSQUERA[a,b,1], Pierina TORRENS[a], Martina AINESEDER[c] and Sonia BENITEZ[a]
[a] *Health Informatics Department, Hospital Italiano de Buenos Aires, Argentina*
[b] *School of Engineering, Universidad Tecnológica Nacional, Argentina*
[c] *Radiology Department, Hospital Italiano de Buenos Aires, Argentina*

Abstract. The acceptance of artificial intelligence (AI) systems by health professionals is crucial to obtain a positive impact on the diagnosis pathway. We evaluated user satisfaction with an AI system for the automated detection of findings in chest x-rays, after five months of use at the Emergency Department. We collected quantitative and qualitative data to analyze the main aspects of user satisfaction, following the Technology Acceptance Model. We selected the intended users of the system as study participants: radiology residents and emergency physicians. We found that both groups of users shared a high satisfaction with the system's ease of use, while their perception of output quality (i.e., diagnostic performance) differed notably. The perceived usefulness of the application yielded positive evaluations, focusing on its utility to confirm that no findings were omitted, and also presenting distinct patterns across the two groups of users. Our results highlight the importance of clearly differentiating the intended users of AI applications in clinical workflows, to enable the design of specific modifications that better suit their particular needs. This study confirmed that measuring user acceptance and recognizing the perception that professionals have of the AI system after daily use can provide important insights for future implementations.

Keywords. user satisfaction; artificial intelligence; clinical decision support systems; radiography; chest.

1. Introduction

The acceptance of AI-based systems by health professionals is crucial to obtain a positive impact on the diagnosis pathway [1]. Understanding why specialists accept or reject a technology is necessary to better predict, explain, and increase user acceptance [2]. The evaluation of user satisfaction can help to identify the system's strengths and weaknesses and guide the planning and design of adequate improvements [3,4].

The primary goal of this study was to obtain a preliminary analysis of user satisfaction with an AI-based system for the automated detection of findings in chest x-

[1] Health Informatics Department, Hospital Italiano de Buenos Aires. 4190 Perón St., C1199AAB, Ciudad Autónoma de Buenos Aires, Argentina. Tel (+54) 01149590200 - ext. 5056; E-mail: candelaria.mosquera@hospitalitaliano.org.ar.

rays, named TRx, which was developed and validated at a health center. In this study, we evaluated the TRx application integrated in the Electronic Health Records (EHR) and the Radiology Information System (RIS) of our center. Our objective was to find patterns in perceptions that were common across users, and identify which factors are implied in the positive uptake of an AI-system for medical imaging, stratifying the results by users' specialties.

Figure 1. TRx interface for RIS users. Red captions indicate the main sections.

2. Methods

This was an IRB-approved observational mixed study (N° 6025), conducted at the Emergency Department of a 650-bed university hospital in Buenos Aires between January 1st and May 31th 2021. A total of 7689 chest x-ray studies were processed by TRx during the study period (average of 51 studies per day). Participants were selected as TRx intended users in the emergency setting: radiology residents and emergency physicians. To assess user satisfaction we evaluated the four factors of the Technology Acceptance Model [7]:

- **Actual system use**: the degree to which a person uses the technology.
- **Perceived usefulness**: the degree to which a person believes that using a particular system would enhance his or her job performance.
- **Perceived ease of use:** the degree to which a person believes that using a particular system would be free from effort.
- **Output quality**: the perceived correctness of the application´s prediction.

TRx is an AI application that assists users in chest x-ray interpretation. It combines four deep learning models that were trained for the detection of four critical findings: pneumothorax, rib fracture, pleural effusion and lung opacities [5]. In the interface, user feedback can be optionally completed at the time of image evaluation. TRx has two different questionnaires to address two distinct intended uses.

(1) When accessed through the RIS (radiology specialists), the questionnaire is focused on diagnostic performance: it consists of a four-point scale on the level of discrepancy with TRx's diagnosis, which was already used for evaluation of the radiology residents' preliminary reports [6] (Fig. 1).

(2) When accessed through the EHR (emergency physician or other specialties), this consists of a five-point Likert scale on the level of usefulness for their work.

3. Results

Regarding actual system use, we retrieved quantitative data on system use from TRx's database, which records the number of times someone accessed the application and found that the interface was accessed in 15.4% of studies (n=1186), with an average of 8 accesses per day. To estimate output quality, we considered the radiologists' feedback: in RIS questionnaires, the proportion of agreement varies greatly among images where TRx detected findings and those where it detected no findings: 90% and 34% respectively (Fig. 2b). Regarding perceived usefulness, we observed that 60% of answers classified TRx with a good utility level in EHR questionnaires (Fig. 2a).

Perceived ease of use was measured through a validated survey using the System Usability Scale (SUS), a widely adopted method for assessing user experience [8,9] (Fig. 3). The survey was answered by 13 professionals: 62% from the Radiology Department and 38% from the Emergency Department. It showed that most users agree that TRx is comfortable and easy to understand, requiring no technical support and no special training to start using it. The greater variation in questions about confidence and consistency suggests these might have been understood as referring to diagnostic performance.

Qualitative analysis was performed by interviewing participants who volunteered, following a structured list of questions. Six physicians were interviewed: three emergency physicians and three radiology residents.

Actual system use: In general, all interviewed participants agreed they would continue using this tool and they would recommend it. An emergency physician noted that she often evaluates X-ray images directly in the portable equipment at the patient's bedside instead of using the EHR, so TRx is not used in those cases. Radiology residents used TRx through the RIS during their X-ray training period, analyzing about 20-30 chest x-rays per day.

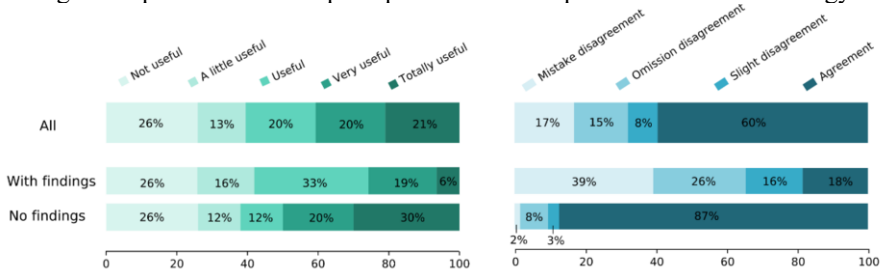
Perceived usefulness: all participants agreed they find the system useful for their daily work, and stated they look at the original image first, to make their own diagnosis, and only after that they look at TRx output as a second opinion. We identified two common TRx uses mentioned as helpful by all interviewed emergency physicians: to confirm a finding has not been omitted, and as an alert of a critical finding, particularly when the patient shows no suggestive clinical signs. Radiology residents also agreed on the utility as diagnosis confirmation, particularly in images with no findings. Additionally, they mentioned TRx's utility to reduce human errors associated with fatigue or rush caused by work overload.

Perceived ease of use: all participants rated the system usability as very good, describing the interface as easy to understand and practical. Suggested improvements included adding tags to identify specific findings in the heatmap and providing visualization tools on the heatmap, such as zoom and scroll.

Output quality: emergency physicians perceived a high output quality, but admitted it might be due to limited use. When comparing it to their own diagnosis, they found they agreed with TRx almost always. However, radiology residents found that diagnostic performance varied greatly across radiological findings. They all concur that the system is especially good at detecting pneumothorax and pleural effusion, and also experienced a good agreement with TRx in normal images. However, they perceived a poor performance for lung opacities, with many false positives. In addition, they noticed a decreased performance in chest x-rays with poor acquisition technique, for example studies from patients who were lying in bed.

4. Discussion

In this work we present the preliminary results of a study on user satisfaction with an AI application for chest x-ray diagnosis implemented in real clinical practice. We found general patterns in users' perceptions on four aspects from the Technology



(a) (b)

Figure 2. User feedback collected from 401 answers. (a) EHR answers. (b) RIS answers

Acceptance Model. The best perceived aspect was the ease of use, which was consistently remarked as very good both in the survey and interviews. Perceived usefulness focused on using TRx to confirm no omission of findings, while emergency physicians also mentioned its utility as an early alert on potential critical findings and radiology residents expressed that TRx guidance could help reduce human diagnostic errors.

A difference among specialties was also observed in the perception of output quality, as physicians generally expressed that TRx has good detection accuracy, while radiology residents provided a detailed judgement over different pathologies, which is aligned with the quantitative results. This confirms that the expected utility of TRx is different in the Emergency and Radiology Departments. This should be considered when adjusting TRx implementation, adapting the system to suit the intended use in the EHR and the RIS respectively.

This study confirmed that output quality is a decisive factor in user satisfaction. A system with good diagnostic performance is the first step to build users' trust in its output, which is certainly required to increase use during daily practice. In particular, results showed that an essential step in TRx future versions should be increasing the performance for lung opacities. We also confirmed that system use is improved by a suitable placement of the application throughout clinical workflows, which was better achieved for the radiology workflow than the emergency one.

The main limitation of this report is its small sample size, which impedes quantitative comparisons with precise statistical results. However, our data shows clear trends and interesting patterns that may guide further efforts in the development and clinical implementation of AI-based diagnostic tools. Future work includes collecting new participants to increase the significance of results.

The main strength of this study is that it reports on data from a real clinical implementation of AI in medical imaging diagnosis. User satisfaction with health AI applications has not been studied widely, and few prior works report on this topic [10-12]. Future research should focus on identifying bottlenecks and barriers that are met by different groups of users when using the system, to allow relevant modifications that could actually improve the system’s clinical utility in real healthcare settings [13].

5. Conclusion

This study of user satisfaction with TRx application yielded positive evaluations from participants. Emergency physicians and radiology residents shared common patterns regarding their perception of TRx's usability, while differing on output quality and usefulness for their work. These results highlight the importance of clearly

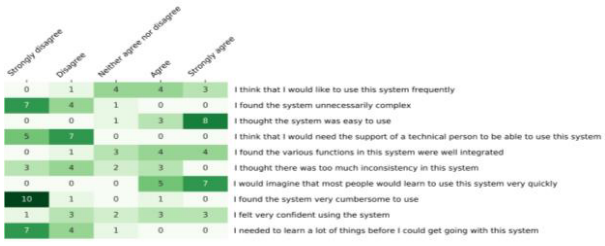


Figure 3. The ten questions on System Usability Scale and the count of survey answers.

differentiating the intended users of AI applications in clinical workflows, to allow a separate analysis of their satisfaction and to adapt specific designs that meet their needs. Measuring user acceptance and recognizing the perception that professionals have of the AI system after daily use can provide important insights for future implementations.

References

[1] He J, Baxter SL, Xu J, Xu J, Zhou X, Zhang K. The practical implementation of artificial intelligence technologies in medicine. Nature medicine. 2019;25(1):30-36.
[2] Høstgaard AM, Bertelsen P, Nøør C. Methods to identify, study and understand end-user participation in HIT development. BMC medical informatics and decision making. 2011;11(1):1-11.
[3] Akhloufi H, et al. A usability study to improve a clinical decision support system for the prescription of antibiotic drugs. PloS one. 2019;14(9):e0223073.
[4] Kastner M, Lottridge D, Marquez C, Newton D, Straus SE. Usability evaluation of a clinical decision support tool for osteoporosis disease management. Implementation Science. 2010;5(1):1-12.
[5] Mosquera C, et al. Chest x-ray automated triage: A semiologic approach designed for clinical implementation, exploiting different types of labels through a combination of four Deep Learning architectures. Computer Methods and Programs in Biomedicine. 2021;206:106130.
[6] Diaz FN, Ulla M. Validation of an informatics tool to assess resident's progress in developing reporting skills. Insights into imaging, 2019;10(1):1-10.
[7] Holden RJ, Karsh BT. The technology acceptance model: its past and its future in health care. Journal of biomedical informatics, 2010;43(1):159-172.
[8] Bangor A, Kortum PT, Miller JT. An empirical evaluation of the system usability scale. Intl. Journal of Human–Computer Interaction, 2008;24(6):574-594.
[9] Laugwitz B, et al. Construction and evaluation of a user experience questionnaire. In Symposium of the Austrian HCI and usability engineering group. Springer, Berlin, Heidelberg. pp. 63-76.
[10] Kim DW, Jang HY, Kim KW, Shin Y, Park SH. Design characteristics of studies reporting the performance of artificial intelligence algorithms for diagnostic analysis of medical images: results from recently published papers. Korean journal of radiology. 2019;20(3):405.
[11] Beede E, et al. A human-centered evaluation of a deep learning system deployed in clinics for the detection of diabetic retinopathy. In Proceedings of the 2020 CHI conference on human factors in computing systems. p. 1-12.
[12] Jauk S, Kramer D, Avian A, Berghold A, Leodolter W, Schulz S. Technology Acceptance of a Machine Learning Algorithm Predicting Delirium in a Clinical Setting: a Mixed-Methods Study. Journal of medical systems. 2021;45(4):1-8.
[13] Magrabi F, et al. Artificial intelligence in clinical decision support: challenges for evaluating AI and practical implications: a position paper from the IMIA technology assessment & quality development in health Informatics working group and the EFMI working group for assessment of health information systems. Yearbook of medical informatics. 2019;28(1);128.

Challenges of Trustable AI and Added-Value on Health
B. Séroussi et al. (Eds.)

doi:10.3233/SHTI220387

Scaling AI Projects for Radiology – Causes and Consequences

Gunnar ELLINGSEN[a,1], Line SILSAND[b], Gro Hilde SEVERINSEN[b] and Line LINSTAD[b]
[a] *UIT – The Arctic University of Norway, Tromsø*
[b] *Norwegian Centre for E-health Research, University Hospital of North Norway, Tromsø*

Abstract. Artificial intelligence (AI) for radiology has the potential to handle an ever-increasing volume of imaging examinations. However, the implementation of AI for clinical practice has not lived up to expectations. We suggest that a key problem with AI projects in radiology is that high expectations associated with new and unproven AI technology tend to scale the projects in ways that challenge their anchoring in local practice and their initial purpose of serving local needs. Empirically, we focus on the procurement of an AI solution for radiology practice at a large health trust in Norway where it was intended that AI technology would be used to process the screening of images more effectively. Theoretically, we draw on the information infrastructure literature, which is concerned with scaling innovative technologies from local settings, with a limited number of users, to broad-use contexts with many users.

Keywords. Artificial intelligence, radiology, socio-technical, scaling

1. Introduction

AI for radiology has the potential to handle an ever-increasing volume of imaging examinations and thus be a countermeasure against the lack of human radiological resources [1]. AI can support radiologists in many of their core responsibilities, such as scheduling examinations, prioritizing images according to severity, and interpreting images [2]. However, the implementation of AI in clinical practice has not lived up to expectations, as only a few solutions have made it into actual use [2-4]. Existing studies have attributed this mainly to limitations in the technology [5] while organizational issues have basically been ignored [3]. In this paper, we suggest that a key problem with AI projects in radiology is that high expectations associated with new and unproven AI technology tend to scale the projects in ways that challenge their anchoring in local practice and their initial purpose of serving local needs. We don't necessarily attribute this to the project per se, but rather see the scaling as the effect of a combination of internal and external factors. We ask the following research question: What causes scaling of AI projects in radiology, how does it happen, and what are the consequences? Empirically, we focus on the procurement of an AI solution for radiology practice at a large health trust (denoted Health Trust N) in Norway. The trust is one of the largest in

[1] Corresponding author, Gunnar Ellingsen, UIT – The Arctic University of Norway, research group SOF NORD, 9038 Tromsø, Norway; E-mail: gunnar.elllingsen@uit.no

Norway, with about 9,800 employees and is responsible for providing specialist health services for about 500,000 people. The motivation for Health Trust N's ambitions of implementing AI solutions is caused by a steady increase in labor-intensive imaging examinations, estimated at 5–10% per year. Therefore, the trust intended to use AI technology to process the screening of images more effectively. Theoretically, this paper draws on the socio-technical literature, meaning that technology must always be understood in the organizational context in which it is implemented and used. Specifically, we draw on the information infrastructure literature, which is concerned with scaling innovative technologies from local settings, with a limited number of users, to broad-use contexts with many users [6]. This concern is sometimes referred to as the "bootstrap problem." In this regard, it is crucial to understand the organizational factors that drive the scaling process and thereby assess how sustainable the scaling process is.

2. Methods

This study is based on an interpretive research approach [7,8], in which reality is socially constructed among participants. We aim to see the causes for scaling the AI project from the viewpoints of different stakeholders while also considering the broader context. Data are collected through approximately hour-long semi-structured interviews with 20 informants, whereof five are radiologists, seven are hospital managers and eight are managers in collaborating organizations. The data analysis is based on a hermeneutic approach, wherein we understand a complex whole from preconceptions about the meanings of its parts and their interrelationships [7]. We discussed the empirical data thoroughly within the research project, as well as with members of the radiology project, to get a balanced picture of the process. In the interviews, we also looked for recurrent themes and patterns, which we explored further in the interviews that followed. All interviews were audio-recorded and transcribed for analysis.

3. Results

In 2019 Health Trust N applied for innovation funds from its regional health authority to explore the possibilities of implementing an AI solution. The Trust was granted a budget of 1.7 MNOK. Apart from this, the project was expected to use internal human resources, and, most importantly, the future key users: the radiologists from the radiology department. In early 2020, the procurement project started the first phase, preparing the procurement of a test solution. This included clarifying the clinical needs, an associated requirements specification and a strategy for evaluating the solution in clinical practice. The project concluded that AI should be used for the following examinations:

- CT (thorax for lung nodules, pulmonary embolism, and lung metastases, which can make the examination process more efficient and improve survival rates from lung cancer.
- MR caput for multiple sclerosis follow-up is an MR scan of the brain, which relieves the radiologists from a time-consuming examination.
- Conventional X-ray for skeletal X-ray and chest X-ray, which can provide faster diagnostics and sift out irrelevant findings.

The procurement process is aligned with EU regulations for public tender acquisitions and conforms to the principle of competitive dialogue. However, after starting out relatively modestly, the project has gradually scaled in ambition, purpose and outlook. In the following, we elaborate on three different ways the scaling is happening.

3.1. From a local to a national scope

First, while there are several research initiatives related to the development of algorithms for diagnostic imaging in Norway, none of these have been implemented in clinical practice. This paradox is of national concern and therefore the Norwegian Directorate of Health has established a national project named "Better Use of Artificial Intelligence" to facilitate the process towards realizing the benefits of AI in clinical practice. Based on Health Trust N's initiative, the Norwegian Directorate of Health, through its regional health authority, commissioned it to explore how the optimal procurement process of a commercial ready-to-use AI solution for clinical practice could be done most efficiently. This positioned the AI project at Health Trust N as the national pilot for such a purpose. One of the project members said: "We try as best we can to keep the regional health authority up to speed on how we are doing things, so that the next trust in line to implement this can do so better and faster."

Second, several other health trusts became interested in Health Trust N's initiative. As a result, eight of these trusts have joined the framework deal that Health Trust N is negotiating and plan to acquire the same solution if this project is successful. However, the various health trusts all have different Radiology Information Systems (RIS) and Picture Archiving Systems (PACS). This is expected to increase the complexity of integration and interoperability because these systems need to communicate with the AI solutions. The heterogeneous portfolios of RIS/PACS in the health trusts make it more urgent for the project to find a solution that will work well for everyone. One of the project members has asked rhetorically: "There is really mounting pressure in the project … what happens if we can't come up with a solution that works for everybody?" Still, the project perceives the common framework deal for several trusts as a win-win situation, and a risk worth taking because it will be beneficial for the Norwegian healthcare sector. Acquiring an AI solution as part of the same framework agreement therefore means that the trusts can put an AI solution into real use much faster than they would otherwise have been able to.

3.2. From algorithm to platform

From early on, the AI procurement project planned to procure a commercially available CE-marked solution tested in another European country to limit the need to validate the algorithms. In addition, the requirement was that the algorithms would be static, meaning that they should not be learning by themselves, that is, by continuous feedback from datasets in use. In August 2021, the project started the procurement process with five international vendors. One vendor offered single algorithms and the other four offered platform solutions. The vendors offered a variety of AI solutions, all conforming to the six different examinations. The platform vendors offered some of the same AI applications developed by third-party vendors, some apps from different third-party vendors, and some apps developed in-house. The vendor offering only single algorithms had developed these in-house. The AI solutions offered by the different vendors ranged

from immature products with weak documentation, both internally and in peer-reviewed publications, to established and mature products that were well documented. The variety of maturity and documentation made it difficult for the members of the project to know which solutions to choose and which one would fit a particular clinical use in a specific organizational context. During the procurement process, the members of the project came to an understanding that choosing applications from different vendors for each of the six use cases would involve more complex technical integration and business processes for the health trust, compared to choosing a platform solution with only one point of integration and only one vendor to address. By choosing a platform vendor, the AI project gained access to several AI applications through a single point of integration to the platform residing in the cloud. If an application did not meet expectations, then it would be easy to replace it with another app from the platform menu. The platform vendor would take care of all legal, ethical and practical collaboration with third-party vendors. One of the project members expressed the expectations of such a platform like this: "We believe that there are some areas where you can reap the benefits straight away, but that the potential is very much greater in the future. And based on this we chose to go for a platform solution."

3.3. From one to several application areas

The project started out rather narrowly, focusing on solving Health Trust N's problem of the lack of radiology resources. However, increasingly, the project is exploring other possible use areas. First, using radiology-related AI in emergency rooms may deal with long waiting lists. AI may then offer preliminary results that enable the triaging of patients. Patients with a low probability of injury or fracture can then be sent home, pending a radiologist's careful examination of the images and a final assessment. The project members emphasize that such solutions have already been implemented at many health institutions in Europe. This could potentially contribute to the more efficient use of resources in the specialist health service and could streamline the operation of the emergency room. Second, AI may potentially improve the practice of referral by general practitioners (GPs) to hospitals by ensuring that the patient's need is described accurately. For example, AI can improve the quality of the referral to the X-ray department so that the correct examination is chosen for the patient's problem. Not all conditions need such advanced examinations as CT and MRI. An MRI examination is quite expensive, plus there are long waiting lists of patients who need this examination. A machine learning algorithm can support the requisition process by assessing the usefulness of certain examinations for given conditions and making this information available to the GP. Third, based on a recent inquiry by the Norwegian Cancer Registry, the AI project is considering how the platform solution can support the Registry's research project on how AI algorithms can make breast cancer screening more efficient. In the Cancer Registry's research project, they have faced quite big challenges in implementing algorithms for testing. In this regard, Health Trust N's platform solution may be a way forward, since it is supposed to be quite easy to add new algorithms.

4. Concluding Discussion

A key concern in the information infrastructure literature is how technology is implemented into real use and how this technology scales from one setting to the next.

This concern is called the "bootstrap problem" since the user community is initially non-existent or very small [6]. One strategy to deal with this is to start with a small and delimited technology, which the targeted users find useful. After the technology is established as a working solution, new user groups can be recruited, and new functionality can gradually be added to complement what already exists, and thus create more value for the users. In this way, scaling takes place gradually and always with a foothold in real use. Also, in our study, the AI project at Health Trust N started quite narrowly with limited ambition, focusing on solving Health Trust N's problem of the lack of radiology resources. The initial idea was to explore the possibilities of implementing commercial ready-to-use AI solutions conforming to the local needs of Health Trust N. However, increasingly, the project is facing scaling challenges as a result of the involvement of many more stakeholders on different levels and expectations of new application areas. A major challenge is that the scaling happens before there is any real use, which in turn can make it difficult to maintain the interest of the original users – the radiologists [6]. While the project is no doubt trying to keep the interest of the radiologists intact, there are now many other stakeholders that expect to be listened to on both the local and the regional levels. The project must balance these interests carefully, which ultimately means that everybody must give and take, which may cause delays and unforeseen results (at least for some). To face some of the challenges associated with scaling, the project has invested in a platform solution. On the one hand, this seems wise, given that the project is in the situation where it needs to align the expectations of different stakeholders. On the other hand, a platform approach also represents a scaling from Health Trust N's initial ambition of simply requiring a ready-made algorithm. While perhaps unintended, the consequence is that both new and old stakeholders may see potential for new application areas, which in turn may further scale the ambitions and scope. An illustration is how the Norwegian Cancer Registry sees great potential in the platform. Also, internally, the project considers new application areas in the emergency room and in the GPs' practices. While these represent long-term prospects, they nonetheless illustrate how easily scaling may happen for the much-applauded technology.

References

[1] Oren O, Gersh, BJ, Bhatt, DL. Artificial intelligence in medical imaging: Switching from radiographic pathological data to clinically meaningful endpoints. The Lancet Digital Health. 2020; 2(9):e486–e488.
[2] Syed AB, Zoga AC. Artificial intelligence in radiology: Current technology and future directions. Seminars in Musculoskeletal Radiology. 2018; 22(5):540–545.
[3] Strohm L, Hehakaya C, Ranschaert ER, Boon WPC, Moors EHM. Implementation of artificial intelligence (AI) applications in radiology: Hindering and facilitating factors. European Radiology. 2020; 30(10):5525–5532.
[4] van Leeuwen KG, Schalekamp S, Rutten MJCM, van Ginneken B, de Rooij M. Artificial intelligence in radiology: 100 commercially available products and their scientific evidence. European Radiology. 2021; 31(6): 3797–3804.
[5] Sogani JAB, Dreyer K, McGinty G. Artificial intelligence in radiology: The ecosystem essential to improving patient care. Clinical Imaging. 2019; 59:8–11.
[6] Aanestad M, Jensen TB. Building nation-wide information infrastructures in healthcare through modular implementation strategies. The Journal of Strategic Information Systems. 2011; 20(2): 161–176.
[7] Klein H, Myers, M. A set of principles for conducting and evaluating interpretive field studies in information systems. MIS Quarterly. 1999; 23: 67-94.
[8] Walsham G. Interpretive case studies in IS research. Nature and method. European Journal of Information Systems. 1995; 4: 74-81.

Challenges of Trustable AI and Added-Value on Health
B. Séroussi et al. (Eds.)

doi:10.3233/SHTI220388

ECG Classification Using Combination of Linear and Non-Linear Features with Neural Network

Tetiana BILOBORODOVA[a], Inna SKARGA-BANDUROVA[b,1], Illia SKARHA-BANDUROV[c], Yelyzaveta YEVSIEIEVA[d] and Oleh BILOBORODOV[e]
[a] *G.E. Pukhov Institute for Modelling in Energy Engineering, Ukraine*
[b] *Oxford Brookes University, United Kingdom*
[c] *Luhansk State Medical University, Ukraine*
[d] *School of Medicine, V.N. Karazin Kharkiv National University, Ukraine*
[e] *Central Research Institute of Armament and Military Equipment of Armed Forces of Ukraine, Ukraine*

Abstract. In this paper, we present an approach to improve the accuracy and reliability of ECG classification. The proposed method combines features analysis of linear and non-linear ECG dynamics. Non-linear features are represented by complexity measures of assessment of ordinal network non-stationarity. We describe the basic concept of ECG partitioning and provide an experiment on PQRST complex data. The results demonstrate that the proposed technique effectively detects abnormalities via automatic feature extraction and improves the state-of-the-art detection performance on one of the standard collections of heartbeat signals, the ECG5000 dataset.

Keywords. ECG, neural network, ordinal partition network, conditional permutation entropy, global node entropy

1. Introduction

An electrocardiogram (ECG) is a standard test to evaluate the heart condition owing to the time change in the total electrical potential. Diagnosis of heart diseases by ECG is based on a morphological analysis of the amplitude-time parameters of the PQRST complex. An ECG is accurate at diagnosing many heart diseases, although it does not always pick up every heart problem. Moreover, an abnormal ECG is highly diverse and can mean even a normal variation of a heart rhythm, not affecting the health. The traditional approaches to differentiate between norm and pathology mainly rely on time series or frequency domain data. The reliable models can be achieved using the nonlinear state-space dynamics analysis that display system behavior in n-dimensional space. Although the nonlinear methods have shown promising results for ECG signal analysis, they can be extended to improve the performance of classification algorithms. For this purpose, the research community is witnessing a rising interest in applying neural

[1] Corresponding Author, Inna Skarga-Bandurova; E-mail: skarga_bandurova@ukr.net.

networks that have shown promising results. The study [1] proposed neural network architecture for classifying PQRST segments of ECG complexes to diagnose cardiovascular diseases. The authors of [2] presented a deep neural network model trained on a dataset of 2,322,513 ECG records to detect six cardiac pathologies.

The further improvement of ECG classification can be achieved through combination neural networks with complementary approaches. The study [3] proposes a method based on the CNN for heartbeat classification using a dual-beat ECG coupling matrix that shows an adequate representation of both heartbeat morphology and rhythm. The use of time-frequency features in addition to neural networks demonstrates a high quality of ECG classification [4, 5]. However, as mentioned, many diagnostic features do not always provide the necessary reliability of diagnostic results and need further exploration [6]. In this concern, we refer to the analysis of the nonlinear ECG dynamics and the ordinal network [7] as an appropriate tool that provides valuable information about cardiac activity changes. We analysed various measures of consistency used to compare time series and distinguish between regular (periodic), chaotic, and random behaviour [8]. The main types of constant measures are entropies, fractal dimensions, and Lyapunov exponents. There are different types of entropy measures applied for ECG analysis, Renyi phase permutation entropy [9], Lempel-Ziv complexity and Lyapunov exponent [10], Kolmogorov-Sinai entropy [11]. Some recent studies use permutation entropy (PE), conditional permutation entropy (CPE), global network entropy (GNE) [7]. The paper [12] demonstrated a change in the degree of the ECG complexity in the event of pathology. All these criteria to some extent characterise the randomness of the mechanisms of cardiac activity and can be used to increase the reliability of assessing the nature of the cardiac activity. This paper proposes a new methodology for analysing ECG time series data, considering linear and nonlinear dynamics. The methodology includes the application of neural networks (NNs) in combination with ECG ordinal network (ON) complexity estimation to classify cardiac activity.

2. Materials and Methods

Our hypothesis is based on the assumption that the ON improves the classification accuracy of NNs and, thus, provides an honestly significant difference in the dynamics deviations of the PQRST cycle. The proposed methodology includes combination of NN and ON analysis, in particular the calculation of complexity measures, to ECG time series data. The mapping the time series in ON is based on its ordinal partition, having regard to sequential patterns of PQRST ECG complexes [7]. Ordinal partitioning is parameterized by the dimension of embedding vector m and embedding delay τ.

We denote the ECG time series as a series of equidistant observations $\mathbf{x} = \{x_1, \ldots, x_m, \ldots, x_M\}$, where m is an integer representing the dimension of the embedding vectors of the time series and M is the number of time steps in the time series being split. The embedding vectors $\mathbf{z} = \{z_1, z_2, \ldots, z_{M-(m-1)\tau}\}$ are calculated using patterns of dimension m and delay τ. For ECG, longer embedding vectors can define PQRST complexes, while shorter embedding vectors can define PQRST complex segments associated with functional components: waves or segments of the complex. Each vector $z_i = \{x_1, x_{i+\tau}, x_{i+2\tau}, \ldots, x_{i+(m-1)\tau}\}$ is mapped into a sequence of symbols according to the amplitude of each element in it. If two z_i elements have the same value ($x_i = x_j$), the character is assigned in order of occurrence. The unique set of all character sequences s is represented by a time series $\mathbf{s} = \{s_1, s_2, \ldots, s_{M-(m-1)\tau}\}$, each $s_i \in \mathbf{s}$ being a node to construct an ordinal network.

The proposed methodology is shown in Fig. 1 and includes the following steps: (1) building an ON and estimating randomness using complexity measures such as conditional permutation entropy (CPE) or global node entropy (GNE); (2) combining the obtained score with ECG data and classifying cardiac activity using NN.

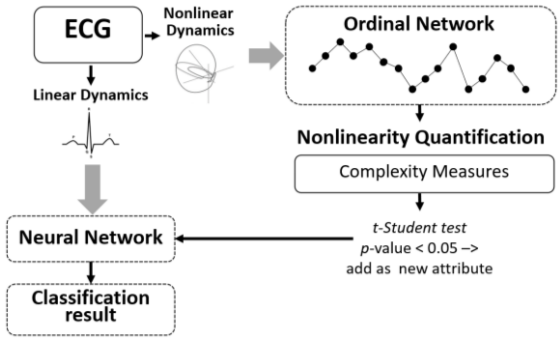


Figure 1. Structure of proposed methodology.

The proposed complexity measures (CPE, GNE) are calculated based on the permutation entropy (PE) of the time series, which corresponds to the Shannon entropy of the corresponding set of ordinal symbols **s**.

$$PE = -\sum_i p_i \log p_i, \tag{1}$$

where p_i denotes the probability mass function $P(S=s_i)$ for $s_i \in \mathbf{s}$ and is estimated by counting the relative occurrence of each symbol in the symbolic dynamics S. PE is discussed in detail in [7, 8]. The stationary distribution is calculated by finding the stochastic matrix P with elements $p_{i,j}$.

Conditional permutation (CPE) was proposed in [13] and is expressed as follows (2)

$$CPE = \sum_i \left(-p_i \sum_j p_{i,j} \log p_{i,j}\right), \tag{2}$$

where p_i is the probability mass function, same as for equation (1), and $p_{i,j}$ is an element of the stochastic matrix P that represents the transition probability from s_i to s_j estimated on the basis of the symbolic dynamics of S.

Global node entropy is calculated based on the elements $p_{i,j}^T$ of the modified stochastic matrix P^T of the ordinal network by averaging over the network based on the stationary distribution (3).

$$GNE = \sum_i p_i \left(-\sum_j p_{i,j}^T \log p_{i,j}^T\right). \tag{3}$$

3. Results

The proposed methodology, in particular CPE, GNE calculation was applied to analyse features of the PQRST segments of the ECG5000 dataset cycle [14]. ECG5000

contains ECG time series examples with 140 time steps. Each sequence corresponds to a single heartbeat, i.e. PQRST cycle from a single patient with congestive heart failure. In this study, two types of PQRST cycles (classes) were used: class 1 – normal (500 examples) and class 2 – R-on-T premature ventricular contraction (500 examples).

To estimate the individual components of the PQRST cycle from the CPE ordinal split, the GNE of each time series was calculated for a short embedding delay. The value of embedding vector is set to $m = 5$, the value of embedding delay is set to $\tau = 8$. The results of statistical analysis are shown in Fig.2.

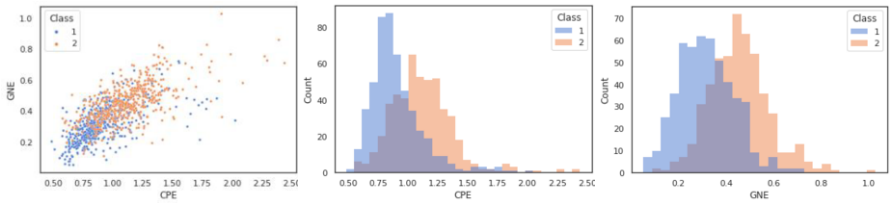


Figure 2. Visualization of the statistical analysis of the CPE and GNE for the test set: (a) the scatter plot of the correlation between the CPE, GNE estimates and data classes, (b) overlaying kernel density plots CPE, (c) kernel density plots GNE.

The scatter plot shows the presence of a certain correlation between the CPE, GNE scores and data classes. By overlaying kernel density plots, we can judge the distribution of score values between classes and infer the presence of a large number of CPE and GNE values that are inherent in a certain class, what gives a certain degree of reliability in their differentiation. Student t-test was applied (p-value lower than 0.05 was considered statistically significant, with a confidence interval of 95%) to test the null hypothesis about the equality of the variances of the obtained estimates of the two classes and obtain a quantitative estimate. For both the CPE and GNE, a p-value of less than 0.05 was obtained, therefore the difference in variances is considered statistically significant, the null hypothesis of equality of variances is rejected. Thus, we start from the assumption that the variances of the CPE, GNE scores for class 1 and class 2 are not equal. The same way, the values of the CPE, GNE estimates were calculated for each PQRST cycle and added as new attributes. For stage 2, the PQRST cycles classification was carried out using a simple implementation of NN - the Sequential model by Keras. NN training was perfomed using the following network hyperparameters: layers activation functions are ReLU and softmax, optimizer is Adam, epochs=100. The models quality results obtained on the test data set are presented in Table 1.

Table 1. Comparative analysis with and without complexity measures.

Attributes	Cross-entropy loss	Accuracy (%)
PQRST complexes attributes	0.0833	99
PQRST complexes attributes +CPE+GNE	0.0213	99.5

Adding the CPE, GNE increases the classification accuracy by 0.5%, and decrease the cross-entropy loss by 0.062. Thus, we can conclude that the classification reliability is increased by using complexity measures as features of non-linear dynamics.

Comparison of our approach to NN-based studies based is presented in Table 2.

Table 2. Comparative analysis with previous research.

Approach	Accuracy (%)
CNN [1]	97.41
CNN+dual-beat ECG coupling matrix [3]	99.1
CNN+ STFT-Based Spectrogram [4]	99.0
CNN+ Continuous Wavelet Transform [5]	98.74
Proposed NN+ON	**99.5**

4. Conclusion

We present a technique that combines the signs of non-linear ECG dynamics with NN to classify ECG which shows 0.4% improvement in accuracy over the best benchmark approach. Hence no-linear dynamics is useful to achieve more regularised predictions and to obtain more reliable classification results for PQRST complexes.

References

[1] Wu M, Lu Y, Yang W, Wong S.Y. A study on arrhythmia via ECG signal classification using the convolutional neural network. Frontiers in computational neuroscience. 2021:106.
[2] Ribeiro AH, Ribeiro MH, Paixão GMM et al. Automatic diagnosis of the 12-lead ECG using a deep neural network. Nat Commun. 2020; 11(1760). https://doi.org/10.1038/s41467-020-15432-4.
[3] Zhai X, Tin C, Automated ECG classification using dual heartbeat coupling based on convolutional neural network. IEEE Access. 2018. 6: 27465-27472.
[4] Huang J, Chen B, Yao B, He W.ECG arrhythmia classification using STFT-based spectrogram and convolutional neural network. IEEE Access. 2019;7:92871-92880.
[5] Wang T, Lu C, Sun Y, Yang M, Liu C, Ou C. Automatic ECG Classification Using Continuous Wavelet Transform and Convolutional Neural Network. Entropy. 2021;23(1):119.
[6] Faynzilberg LS, Mathematical methods for evaluating the usefulness of diagnostic features, Osvita Ukrainy, Kyiv, 2020:152
[7] McCullough M, Small M, Iu HHC, Stemler T. Multiscale ordinal network analysis of human cardiac dynamics. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences. 2017; 375(2096):20160292.
[8] Bandt C, Pompe B. Permutation entropy: a natural complexity measure for time series. Physical Review Letters. 2002;88(17):174102–0. doi:10.1103/physrevlett.88.174102.
[9] Kang H, Zhang X. The influence of parameter selection for Renyi phase permutation entropy on abnormal change detection. IEEE International Conference on Signal, Information and Data Processing (ICSIDP), 2019. p. 1-6.
[10] Sun Q, Wang Q, Ji B, Wu W, Huang W, Wang C. The Cardiodynamicsgram Based Early Detection of Myocardial Ischemia Using the Lempel-Ziv Complexity. IEEE Access; 2020.8:207894–207904.
[11] Scarciglia, A., Catrambone, V., Bonanno C., Valenza, G. Quantifying partition-based Kolmogorov-Sinai Entropy on Heart Rate Variability: a young vs. elderly study. 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC), 2021, pp. 5469-5472, doi: 10.1109/EMBC46164.2021.9630975.
[12] Mayorov, O., Fenchenko V. Increase reliability of bioelectric activity (EEG, ECG and HRV) deterministic chaos researches by the nonlinear analysis methods. J. Klin.inform.telemed; ISSN 1812-7231; 5 (6), 10 - 17.
[13] Unakafov AM, Keller K. Conditional entropy of ordinal patterns. Physica D: Nonlinear Phenom. 2014;269:94–102.
[14] Chen Y, Hao Y, Rakthanmanon T, Zakaria J, Hu B, Keogh E. A general framework for never-ending learning from time series streams. Data mining and knowledge discovery, 2015;29(6):1622-1664.

Challenges of Trustable AI and Added-Value on Health
B. Séroussi et al. (Eds.)

doi:10.3233/SHTI220389

Dataset Comparison Tool: Utility and Privacy

João COUTINHO-ALMEIDA[a,c,1] , Ricardo João CRUZ-CORREIA[b] and Pedro Pereira RODRIGUES[b]
[a] *CINTESIS - Centre for Health Technologies and Services Research, University of Porto, Portugal*
[b] *MEDCIDS – Faculty of Medicine of University of Porto, Portugal*
[c] *Health Data Science PhD Program, Faculty of Medicine of the University of Porto, Portugal*

Abstract. Synthetic data has been more and more used in the last few years. While its applications are various, measuring its utility and privacy is seldom an easy task. Since there are different methods of evaluating these issues, which are dependent on data types, use cases and purpose, a generic method for evaluating utility and privacy does not exist at the moment. So, we introduced a compilation of the most recent methods for evaluating privacy and utility into a single executable in order to create a report of the similarities and potential privacy breaches between two datasets, whether it is related to synthetic or not. We catalogued 24 different methods, from qualitative to quantitative, column-wise or table-wise evaluations. We hope this resource can help scientists and industries get a better grasp of the synthetic data they have and produce more easily and a better basis to create a new, more broad method for evaluating dataset similarities.

Keywords. Synthetic data, utility evaluation, privacy evaluation

1. Introduction

Synthetic data can be defined as data that has no connection with a real-world phenomena or event. It was not originated from a process in the real world, but rather a synthetic one. The idea is that synthetic data can have similar properties with the real data, without needing to have an independent process for its generation.

Synthetic data has been used over the years for several usages, but in healthcare is still not very used. However, this scenario seems to be changing. It can be used for several use cases namely [1]; i) Software testing, ii) educational purposes, iii) machine learning, iv) regulatory, v) retention, vi) secondary and vii) enhanced privacy.

Software testing relates to using synthetic data to create use cases for software testing. This can be used for the development or pre-production stages for example. Often the data needed is not available on-demand and a synthetic generator of reliable data could be useful. Educational purposes relate to, at least, two different scenarios. One is for onboarding of employees [1], other is related to healthcare students for using health information systems and creating mechanisms for providing reliable data on-demand.

[1] Corresponding Author, João Coutinho-Almeida; E-mail: joaofcalmeida@outlook.com

Machine-learning is one of the areas where synthetic data has more widespread usage, where data augmentation through data synthesis is already common. It can be used for class-imbalance, sample-size boosting or machine-learning algorithms testing. Regulatory purposes could be important as well, with the rise of Artificial Intelligence (AI) as medical device systems and synthetic data could be used to properly evaluate these systems under controlled environments. Retention can be an important case for synthetic data as well, since personal data must not be kept more than it would be necessary. Synthetic data generators can be of use, where the original data can be deleted and a generator kept for further usage, given that the privacy mechanisms are properly employed. Secondary uses relate to using synthetic data to share data with academia or industry. Simulacrum [2] is a nice example of how the NHS uses these mechanisms to help scientists get a better grasp of data before having to fill documentation to query the real data. The same occurs for Integraal Kankercentrum Nederland (IKNL), which has a synthetic version of the cancer registry for scientific purposes [3] and the Medical and Healthcare products Regulatory Agency (MHRA) that uses synthetic data as well for its CPRD real-word evidence [4].

Finally, an aspect that is underlying to all these applications is the promise that synthetic data can be used to improve privacy. Even though specially tweaked data generators can be used to create more privacy-aware datasets, it will be inherently always at the cost of some utility [5]. So, even though synthetic data it is not the silver bullet as primarily thought, synthetic data generation can be undeniably used to help create more private data for all the use cases seen above at the cost of its utility. As for proper methods of evaluating security and utility, are, for now, open research questions. At the present time, it is still complicated to properly assess the utility of the generated data. We have qualitative and quantitative methods. Qualitative methods are related to plots, quantitative are related to some value that defines a evaluation metric. These quantitative metrics can be applied to equal columns from each data set, pair of columns from each dataset or applied over the whole datasets. As for privacy metrics, the metrics rely on duplicates. Full duplicates or membership inference related.

So, in this paper, we developed a data pipeline for data analysis in order to create a report for providing several metrics for data utility and privacy.

2. Methods

The pipeline relies on python and latex for creating the document. It relies also on several packages that implemented methods for evaluating data, namely *scipy* [6], *sdmetrics* [7] and *scikit-learn* [8] and *mlxtend* [9]. Its basis is related to uploading 2 datasets, and a report in pdf is produced. Being that is based on programmatic code, it can be easily converted into API.

The report has a section for dataset description, columns removed due to high-null and brief variable overview. Then a null comparison is done by column and dataset. Following this is the utility subsection. Firstly, by visual methodologies: heat maps for the correlation, bar plots for categorical, density plots for continuous and a pair plot for an overview. As for the quantitative utility evaluation, we divided it column-wise, pairwise and table-wise. The first comprehends the *Kolmogorov–Smirnov* test for continuous and chi-squared test for categorical variables. Distance metrics were also applied to categorical columns. First, they are transformed into distributions and then distance metrics are applied. The results are a descriptive overview of the distance

metrics, having minimum value, average, max value and standard deviation. The distance metrics selected were *Jensen-Shannon Divergence*, *Wasserstein distance*, *Kullback–Leibler divergence* and entropy.

As for pair-wise metrics, we used a discrete and continuous *Kullblack-Leibler divergence*. In this, two pairs of continuous columns are compared using *Kullback–Leibler divergence*. For this, they are put into bins for further application. The same is applied to categorical columns without binning. As for table-wise metrics, first, we used likelihood metrics. We fitted several Gaussian Mixture models or Bayesian network models to the real data and then calculate the likelihood of the synthetic data belonging to the same distribution. The metrics are likelihood for Gaussian mixture and Bayesian models and log-likelihood for the Bayesian model as well.

Then we used machine-learning models (linear regression and decision trees) to assess how similar models behave on both datasets. First, we tested on the same dataset in order to compare evaluation metrics. Then we cross-tested, meaning that the training set was drawn from one dataset and the test set was drawn from the second dataset. Finally, data privacy constraints duplicate evaluation, duplicate existence by removal of a single column and a record linkage approach. With the record linkage, we define a record linkage blocking ("age" in the example) and then try to match rows from the synthetic dataset to the real, with varying known attributes. Then matrix, euclidean and cosine distance was also calculated. Even though it is used for privacy evaluation, by definition, we could also use it for utility assessment. For proper assessment, the continuous and categorical variables should be indicated at the start of the code. The metrics are listed in the table 1.

Table 1. Metrics Assessed

Metric	Method	Context
Bar Plot	Visual	Utility
KDE Plot	Visual	Utility
Heat-map	Visual	Utility
Pair-plot	Visual	Utility
KS test	Column-quantitative	Utility
ChiSquared Test	Column-quantitative	Utility
Kullback–Leibler divergence	Column-quantitative	Utility
Jensen-Shannon Divergence	Column-quantitative	Utility
Wasserstein distance	Column-quantitative	Utility
Entropy	Column-quantitative	Utility
DiscKLD	Table-quantitative	Utility
ContinuousKLD	Table-quantitative	Utility
BNLikelihood	Table-quantitative	Utility
BNLogLikelihood	Table-quantitative	Utility
GMLogLikelihood	Table-quantitative	Utility
Same dataset ratio	Table-quantitative	Utility
Support rules	Table-quantitative	Utility
Different dataset validation	Table-quantitative	Utility
Duplicates	Quantitative	Privacy
Duplicate minus 1	Quantitative	Privacy
Record Linkage	Quantitative	Privacy
Matrix distance	Quantitative	Privacy/utility
Cosine distance	Quantitative	Privacy/utility
Euclidean distance	Quantitative	Privacy/utility

3. Results

A trial example of comparing data is available for data in the UCI repository, namely the heart disease dataset [10]. The synthetic data was created by using the synthpop package [11]. The variables evaluated are listed in table 1. The code can be seen in https://github.com/joofio/dataset-comparasion-report. As an example, the images for visual analysis for categorical (figure 1) and continuous variables (figure 2) are presented below.

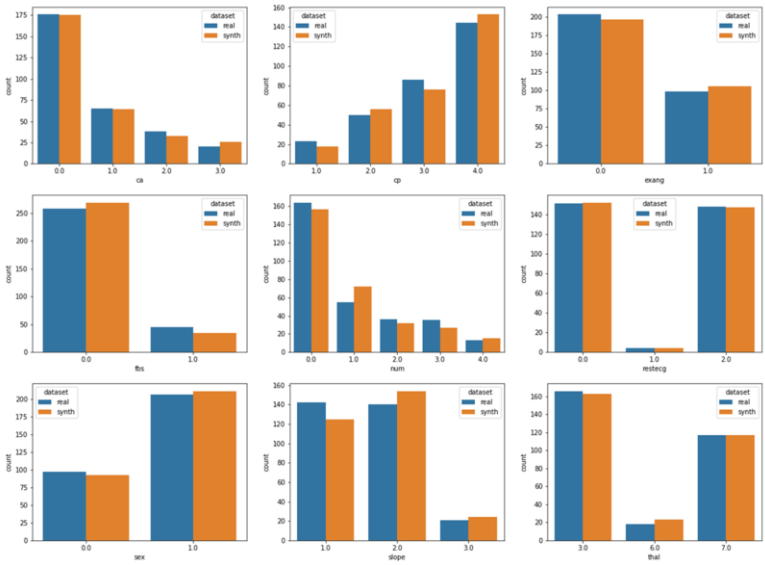


Figure 1. Categorical Variables Plotted

4. Discussion and Conclusions

The compiled evaluation metrics between two datasets are important not only for synthetic vs real datasets evaluation. For example, in distributed learning, where different silos exist, with similar or even equal features, a method for evaluating the similarities can be useful for understanding how the populations are similar between them, trying to shed light on the most similar among them, or different in order to understand the differences in the silos or data acquisition inside them. Furthermore, the differences can be assessed on a more granular level. The column-wise similarities can be different from the inter-columns' similarities and differences between these two metrics can be a metric of interest in itself regarding the quality of the synthetic data and its generator.

With this work, we hope to help institutions and academics getting to access to a benchmark of the datasets provided in order to leverage synthetic data in the healthcare space. Finally, we hope this work helps other researchers reach an evaluation metric that could be a unique and clear response to the question of how similar two datasets are.

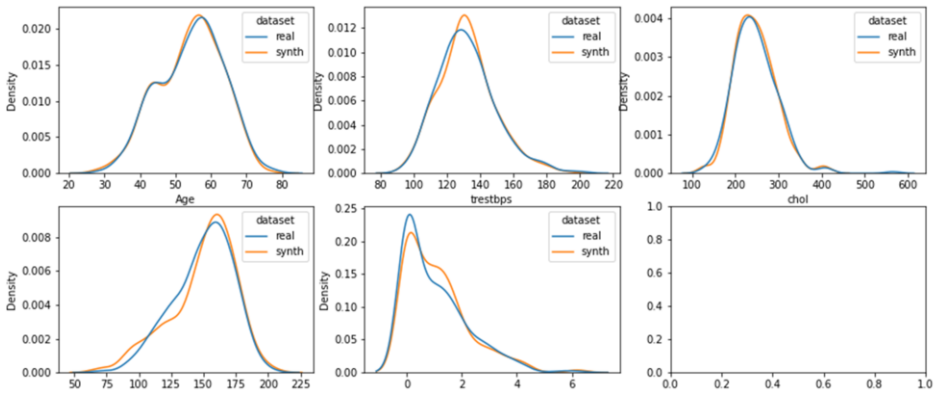


Figure 2. Continuous Variables Plotted

Acknowledgement

This work has been done under the scope of - and funded by - the PhD Program in Health Data Science of the Faculty of Medicine of the University of Porto, Portugal - heads.med.up.pt.

References

[1] James S, Harbron C, Branson J, et al. Synthetic data use: exploring use cases to optimise data utility. Discov Artif Intell. 2021;1:15.
[2] The Simulacrum [Internet]. healthdatainsight.org.uk. 2022 [cited 2022 Jan 21]. Available from: https://healthdatainsight.org.uk/project/the-simulacrum/.
[3] Synthetische dataset NKR beschikbaar voor onderzoekers [Internet]. [cited 2022 Jan 21]. Available from: https://iknl.nl/nieuws/2021/synthetische-data-nkr-beschikbaar-voor-onderzoeker.
[4] Clinical Practice Research Datalink. CPRD cardiovascular disease synthetic dataset [Internet]. Clinical Practice Research Datalink; 2020 [cited 2022 Jan 21]. Available from: https://cprd.com/cprd-cardiovascular-disease-synthetic-dataset.
[5] Stadler T, Oprisanu B, Troncoso C. Synthetic data – A privacy mirage. arXiv. 2020;
[6] Virtanen P, Gommers R, Oliphant TE, et al. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. Nat Methods. 2020;17:261–272.
[7] Patki N, Wedge R, Veeramachaneni K. The Synthetic Data Vault. 2016 IEEE Int Conf Data Sci Adv Anal DSAA [Internet]. Montreal, QC, Canada: IEEE; 2016 [cited 2022 Jan 20]. p. 399–410. Available from: http://ieeexplore.ieee.org/document/7796926/.
[8] Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: Machine Learning in Python. J Mach Learn Res. 2011;12:2825–2830.
[9] Raschka S. MLxtend: Providing machine learning and data science utilities and extensions to Python's scientific computing stack. J Open Source Softw [Internet]. 2018;3. Available from: http://joss.theoj.org/papers/10.21105/joss.00638.
[10] Janosi AMD, Steinbrunn W, Pfisterer M, Detrano R &MD. Heart Disease. 1988.
[11] Nowok B, Raab GM, Dibben C. synthpop: Bespoke Creation of Synthetic Data in R. J Stat Softw. 2016;74:1–26.

Challenges of Trustable AI and Added-Value on Health
B. Séroussi et al. (Eds.)

doi:10.3233/SHTI220390

AP-HP Health Data Space (AHDS) to the Test of the Covid-19 Pandemic

Christel DANIEL[a,b,1], Nicolas PARIS[a], Olivier PIERRE[a], Nicolas GRIFFON[a,b], Stéphane BREANT[a], Nina ORLOVA[a], Patricia SERRE[a], Damien LEPROVOST[a,b], Stéphane DENGLOS[a], Alexandre MOUCHET[a], Julien DUBIEL[a], Rafael GOZLAN[a], Gilles CHATELLIER[a], Romain BEY[a], Marie FRANK[c], Claire HASSEN-KHODJA[d], Marie-France MAMZER[e], Martin HILKA[a] and AP-HP Covid CDW Initiative

[a] *Innovation and Data, IT Department, AP-HP, Paris, France*
[b] *LIMICS, F-75006, Paris, France; Sorbonne Universités, Paris, France*
[c] *Department of Medical Information (DIM), Hôpitaux Universitaires Paris-Saclay, AP-HP Paris, France*
[d] *Direction de la Recherche Clinique et de l'Innovation, Paris, France*
[e] *Université de Paris, AP-HP, Paris, France*

Abstract. Sharing observational and interventional health data within a common data space enables university hospitals to leverage such data for biomedical discovery and moving towards a learning health system. Objective: To describe the AP-HP Health Data Space (AHDS) and the IT services supporting piloting, research, innovation and patient care. Methods: Built on three pillars – governance and ethics, technology and valorization – the AHDS and its major component, the Clinical Data Warehouse (CDW) have been developed since 2015. Results: The AP-HP CDW has been made available at scale to AP-HP both healthcare professionals and public or private partners in January 2017. Supported by an institutional secured and high-performance cloud and an ecosystem of tools, mostly open source, the AHDS integrates a large amount of massive healthcare data collected during care and research activities. As of December 2021, the AHDS operates the electronic data capture for almost +840 clinical trials sponsored by AP-HP, the CDW is enabling the processing of health data from more than 11 million patients and generated +200 secondary data marts from IRB authorized research projects. During the Covid-19 pandemic, AHDS has had to evolve quickly to support administrative professionals and caregivers heavily involved in the reorganization of both patient care and biomedical research. Conclusion: The AP-HP Data Space is a key facilitator for data-driven evidence generation and making the health system more efficient and personalized.

Keywords. Data space, Clinical Data Warehouse, Electronic Health Record, Real-world data, Covid-19 pandemics.

1. Introduction

The analysis of "Real world" data (RWD) has generated important medical discoveries, especially in areas where traditional clinical trials would be unethical or infeasible [1] and represents an opportunity to optimize clinical trials [2,3]. In the era of big data, the

[1] Corresponding Author, Christel DANIEL, Innovation and Data, IT Department, AP-HP, Paris, France; E-mail:christel.daniel@aphp.fr

rise of data science and in particular of new Artificial Intelligence (AI) technologies e.g Machine Learning (ML), RWD is increasingly used to develop innovative digital technologies and new services supporting various activities of health professionals, in research, care and training [4]. RWD and AI should enable healthcare professionals to objectively and continuously measure what is being done, to compare the results of care over time and between health establishments, to question the gaps, to simulate and evaluate the impacts of certain strategic decisions on organizations or the quality of care.

In this context, the objective of this paper is to describe the AP-HP Health Data Space (AHDS) and the IT services that enable the reuse of observational and interventional health data to support clinical and translational research and to move towards a learning health system [5,6].

2. Material and Methods

Assistance Publique - Hôpitaux de Paris (AP-HP) is the largest university hospital in Europe with 39 hospitals (20,500 beds), mainly located in Greater Paris region, conducting more than 3,500 ongoing research projects all sponsors combined. Built on three pillars – **governance and ethics, technology and value proposition** – the AHDS has been developed since 2015 and its major component, the Clinical Data Warehouse (CDW) (https://eds.aphp.fr), made available at scale to health professionals and researchers in January 2017.

Given the importance of the ethical and societal issues raised by the rapid increase of digitalization and data-driven approaches, a **specific governance** has been set up at the strategic and operational levels in order to define the data access and use policy and ensure its implementation guaranteeing patients' rights and privacy.

Supported by an institutional **secured and high-performance cloud and an ecosystem** of tools, mostly open source, the AHDS integrates a large amount of massive healthcare data collected during care and research activities. Initially focused on a core set of data collected through the common Electronic Health Record (EHR)(ORBIS), the Extract-Load-Transform (ETL) process of the AP-HP CDW is now extended to various additional sources including legacy AP-HP applications as well as external sources in order to get a 360° view of patients' health status. Big data such as medical images – and soon genetic sequencing data – are not exhaustively duplicated within the AHDS but copied on demand if needed for any research project.

Beyond a **costly investment** in hardware, software, the implementation of the infrastructure and services supporting the secure and efficient processing of high throughput health data the ADHS also requires the collaboration of experts in sufficient numbers in various domains: clinicians, epidemiologists, IT engineers, biostatisticians and data scientists. Different funding approaches are currently under consideration to sustain the AHDS.

3. Results

The **governance structure of the AHDS** is based on a Steering Board and an Institutional Review Board bringing (IRB00011591) whose members are healthcare professionals, researchers, directors, patient representatives and persons qualified in ethics and that assesses all research projects requiring the use of AP-HP CDW data.

Patients cared at AP-HP are informed individually, via an explicit mention on all medical notes and patient booklets, of both existence and purposes of the CDW as well as the procedures for opting out. The rules of governance as well as the list of the studies carried out or in progress are available on the AHDS web site (https://eds.aphp.fr). In order to guarantee patient privacy, de-identification processes, initially implemented on structured data and extended to medical images and textual documents, are assessed by both the AP- HP Chief Information Security Officer (CISO) and Data Protection Officer (DPO). The **technical pillar** relies on the capacity of the Big Data platform the AHDS offers a set of services supporting data access and use. The AHDS operates large scale clinical trials and cohorts and an increasing number of real-world studies. Most of the interventional data of clinical trials and cohorts sponsored by AP-HP are captured using the institutional electronic data capture (EDC) system (CleanWeb). The recent deployment of the open source EDC system RedCap resulted in new research projects. Observational data are stored in the AP-HP CDW, enabling, as of December 2021, the processing of health data from more than 11 million patients (Table 1). The integration of various data from heterogeneous systems developed in silos, relies on a standardization process based on standard models - such as the OMOP model of the OHDSI project and HL7 FHIR - and reference terminologies – such as ICD10, LOINC, ATC, etc. that contributes to the FAIRification of the data (i.e to make the data Findable, Accessible, Interoperable and Reusable).

Table 1. Content of the AP-HP CDW (as of December 31, 2021). Focus on the research data mart in OMOP format.

Data category	Year of first integration (or expected year)
Demographic data	2017
Vital status (from a national data base)	2021
Clinical pathway and care sites	2019
DRG codes	2017 (ORBIS hospitals), 2022 (all)
Clinical documents	2017 (ORBIS hospitals), 2022 (all)
Vital signs	2017
Consultation and operative schedule	2021
Lab results	2017 (ORBIS hospitals), 2022 (all)
Anatomic pathology precedures	2020
Diagnostic codes, digital slides	2022
Radiological procedures	2021
Medial imagine (DICOM)	2017
Drug prescription/administration	2019 (ORBIS hospitals), 2022 (all)
Care plan	2017
Problems, family and personal history	2022
Covid (coronaOMOP)	2020
Cancer (drug/chemotherapy, prescription,multidisciplinary meeting reports)	2022
Patient reported outcomes (PROMS)	2022
Obstetrics, Infectious diseases, Nephrology, Pediatrics, Rare disease, Emergency care, etc.	
Social history	2023
Exposome : Air pollution, socio-economic data	2022

Between January 1, 2017, and December 31, 2021, 216 research projects have been approved by the IRB. Investigators used the AP-HP cohort builders (i2b2 since January 2017 and now the cohort360 open-source application developed in 2021) to execute direct queries on the CDW and generate secondary data marts. The detailed list of studies is available on the AP-HP CDW web site; their number and nature are described in Table 2. Research projects have significantly increased since 2017 in conjunction with the

involvement of additional data scientist recruited with Covid-dedicated funds or disease-specific donation (e.g., in the domain of cancer). Epidemiological studies performed through the CDW were designed with various purposes: for fine-grained analysis of specific phenotypes, to confirm known disease risk factors or consisted in diagnostic, prognostic or medico-economic studies (table 2). Almost half of the projects were performed using AI/ML approaches particularly in the medical imaging domain. Some of these projects have a clear objective of digital innovation development/assessment (n=61).

Table 2. Number and type of research and innovation projects based on the AP-HP CDW between January 1, 2017 and December 31, 2021

Year	Number of projects	Descriptive	Etiology	Diagnostic	Prognostic	Medico-economical	Digital innovation	Publications
2017	12	3	3	1	4		4	4
2018	26	11	2	5	7	1	13	4
2019	34	17		6	12	1	12	1
2020	95	46	4	11	35	1	18	11
2021	49	29	4	10	10		14	Nd

The also AHDS supports real-world studies implemented at national (French Health Data Hub (HDH)) or international scale (EHDEN or EHR2EDC consortium [7]. Thanks to the Pilote software application, based on the Cognos (IBM) solution, the AP-HP CDW data is also used for strategic and medico-economic management.

Regarding the **valorization pillar**, since the launch of the AHDS in 2015, AP-HP is increasingly investing each year into the development of robust IT solutions to foster collaborative data use by AP-HP both healthcare professionals and public or private partners. The IT staff in charge of the AHDS increased from 5 to +60 persons between 2015 and 2021. In order to meet a growing demand exceeding the capacity to fulfill data requests, funding approaches, such as fee-for-service, are currently under consideration. AP-HP created the AP-HP Research Foundation in 2015 to support biomedical and health research. The return on investment currently consists in publications (>20 articles in 4 years) and grants gained for large-scale national and European projects.

The **impact of the COVID-19 pandemic on the AHDS** has been deep and wide. One month after the admission to the AP-HP of the first patient diagnosed with Covid-19, a specific database was compiled from the CDW. In addition to demographic data, patient pathways, DRG codes (diagnoses and procedures), lab tests, drug prescription, the COVID database includes data of interest automatically extracted from clinical notes using NLP as well as medical images). An information notice specific to the COVID database has been added to the CDW website. The COVID database has been used to produce dashboards informing the crisis unit on a daily basis on the clinical and biological characteristics of patients diagnosed with Covid-19, the evolution of their state of health, and the admission and treating capacities of the hospitals. Evidence-based guidance has been produced for the direction of the AP-HP as well as for policymakers and health authorities on how to mitigate the impact of the outbreak. In the field of research, more than 70 data access requests have been processed by the IRB according to an accelerated procedure. The Covid-19 Research Steering Committee has also prioritized 14 therapeutic trials and 23 other studies. The standardization of data of the COVID database has enabled AP-HP to join the international consortium 4CE (Consortium for Clinical Characterization of Covid19 by EHR (4CE)) whose objective is to rapidly share and analyze aggregated data from hospitalized patients diagnosed with Covid-19 on a large scale in order to better characterize this disease [8].

4. Discussion and Conclusion

In the era of big data and machine learning, the use of RWD, and especially of EHR data will play a growing role in evidence generation in association with randomized clinical trials for promoting scientific research and improving health care delivery in the public interest. The AP-HP Health Data Space (AHDS) is a key facilitator for hospital management, clinical research and digital innovation. The AHDS has been set up with a view to a "rapid and reactive learning system", with the objective of enabling healthcare professionals to access consolidated and enriched health data, regardless of the context – care or research – in which they are collected, and to use these data to improve patient care. The current health crisis highlights the need for healthcare institutions to continue the development and deployment of CDW and Big Data spaces, to strengthen their expertise in data science and to implement efficient data quality monitoring programs. University hospitals are playing and will play a major role in real-world evidence generation. Given the level of funding dedicated to the development IT infrastructures that facilitate high-quality data access and use and the expectations placed on data-driven approaches in biomedical research and health care, there is a need of sound evaluation of the return on public investment in terms of impact of health data reuse infrastructure and projects on care processes, productivity and costs, patient safety, care quality, or health outcomes [10].

References

[1] Eichler H-G, Pignatti F, Schwarzer-Daum B, Hidalgo-Simon A, Eichler I, Arlett P, et al. Randomized Controlled Trials Versus Real World Evidence: Neither Magic Nor Myth. Clin Pharmacol Ther. 2021 May;109(5):1212–8.

[2] Rogers JR, Lee J, Zhou Z, Cheung YK, Hripcsak G, Weng C. Contemporary use of real-world data for clinical trial conduct in the United States: a scoping review. J Am Med Inform Assoc. 2020 Nov 4;28(1):144–54.

[3] Kwakkenbos L, Imran M, McCall SJ, McCord KA, Fröbert O, Hemkens LG, et al. CONSORT extension for the reporting of randomised controlled trials conducted using cohorts and routinely collected data (CONSORT-ROUTINE): checklist with explanation and elaboration. BMJ. 2021 Apr 29;373:n857.

[4] Brnabic A, Hess LM. Systematic literature review of machine learning methods used in the analysis of real-world data for patient-provider decision making. BMC Med Inform Decis Mak. 2021 Feb 15;21(1):54.

[5] Friedman C, Rubin J, Brown J, Buntin M, Corn M, Etheredge L, et al. Toward a science of learning systems: a research agenda for the high-functioning Learning Health System. J Am Med Inform Assoc. 2015 Jan;22(1):43–50.

[6] Richesson RL. Learning health systems, embedded research, and data standards-recommendations for healthcare system leaders. JAMIA Open. 2020 Dec;3(4):488–91.

[7] Griffon N, Pereira H, Djadi-Prat J, García MT, Testoni S, Cariou M, et al. Performances of a Solution to Semi-Automatically Fill eCRF with Data from the Electronic Health Record: Protocol for a Prospective Individual Participant Data Meta-Analysis. Stud Health Technol Inform. 2020 Jun 16;270:367–71.

[8] Brat GA, Weber GM, Gehlenborg N, Avillach P, Palmer NP, Chiovato L, et al. International electronic health record-derived COVID-19 clinical course profiles: the 4CE consortium. NPJ Digit Med. 2020;3:109.

[9] Daniel C, Serre P, Orlova N, Bréant S, Paris N, Griffon N. Initializing a hospital-wide data quality program. The AP-HP experience. Comput Methods Programs Biomed. 2019 Nov;181:104804.

[10] Campion TR, Craven CK, Dorr DA, Knosp BM. Understanding enterprise data warehouses to support clinical and translational research. J Am Med Inform Assoc. 2020 Jul 1;27(9):1352–8.

Challenges of Trustable AI and Added-Value on Health
B. Séroussi et al. (Eds.)

doi:10.3233/SHTI220391

MISeval: A Metric Library for Medical Image Segmentation Evaluation

Dominik MÜLLER[a,b,1], Dennis HARTMANN[a], Philip MEYER[a,b], Florian AUER[a], Iñaki SOTO-REY[b] and Frank KRAMER[a]
[a] *IT-Infrastructure for Translational Medical Research, University of Augsburg, Germany*
[b] *Medical Data Integration Center, Institute for Digital Medicine, University Hospital Augsburg, Germany*

Abstract. Correct performance assessment is crucial for evaluating modern artificial intelligence algorithms in medicine like deep-learning based medical image segmentation models. However, there is no universal metric library in Python for standardized and reproducible evaluation. Thus, we propose our open-source publicly available Python package MISeval: a metric library for Medical Image Segmentation Evaluation. The implemented metrics can be intuitively used and easily integrated into any performance assessment pipeline. The package utilizes modern DevOps strategies to ensure functionality and stability. MISeval is available from PyPI (miseval) and GitHub: https://github.com/frankkramer-lab/miseval.

Keywords. Biomedical image segmentation; Medical Image Analysis, Reproducibility, Evaluation, Open-source framework, Performance assessment

1. Introduction

In the last decade, computer vision analysis based on artificial intelligence methods like deep learning has seen rapid growth in prediction capabilities [1]. This resulted in clinicians striving to integrate computer vision algorithms, like image segmentation, into the medical field. Medical image segmentation (MIS) covers the automated identification and annotation of medically relevant regions of interest (ROI), which can be organs, cell structures, or medical abnormalities like tumors [2]. The idea, especially in radiology and pathology, is to establish these MIS methods in their clinical routine to reduce time-consuming processes and to aid in diagnosis as well as treatment decisions [1]. However, due to the direct impact on medical decisions, the correct evaluation of MIS models is crucial. Nevertheless, recent studies indicated widespread statistical bias in evaluations of MIS models which is also caused by incorrect metric implementation [3]. Furthermore, to our knowledge, there is no universal metric library in Python for standardized and reproducible evaluation. In this work, we propose our open-source publicly available Python package MISeval, which is a metrics library for correct MIS model evaluation. It facilitates an intuitive and fast usage of various popular metrics from literature, as well as ensures implementation functionality and stability.

[1] Corresponding Author, Dominik Müller, IT Infrastructure for Translational Medical Research, Alter Postweg 101, 86159 Augsburg, Germany; E-mail: dominik.mueller@informatik.uni-augsburg.de

2. Methods

The open-source Python module MISeval is a metric library for **M**edical **I**mage **S**egmentation **Eval**uation. The library contains various commonly used metrics for image segmentation, which can be easily imported and instantly used for model performance assessment. MISeval is structured as an API with a central core interface for intuitive usage and is implemented in the programming language Python, which is platform-independent and highly popular for computer vision tasks. This allows simple and fast integration of MISeval in commonly used platforms like Tensorflow, PyTorch, or any NumPy-compatible image segmentation pipeline.

2.1. Metric library

Over the last decades, the MIS literature introduced a large variety of metrics for evaluation. Especially for semantic segmentation, model performance assessment can be quite complex due to the need for scoring pixel classification as well as localization correctness between predicted and annotated segmentation. The MISeval metric library contains popular metrics like Dice Similarity Coefficient (DSC), Intersection-over-Union (IoU), Sensitivity (Sens), Speecificity (Spec), Pixel Accuracy (Acc), AUC, Cohen's Kappa (Kap) and Average Hausdorff Distance (AHD), but also more complex metrics like entropy-based divergence and boundary-based distances. A summary of all metrics in MISeval, can be seen in Table 1.

2.2. Core Interface: Evaluate()

The core of our package is the *evaluate()* function, which acts as a simple and intuitive interface to access and run all implemented metrics. The desired backbone metric for the *evaluate()* function can be defined by passing the name of an already implemented metric or by passing a user-created metric function for uncomplicated integration of custom metrics. Moreover, our core function handles automatically binary as well as multi-class problems. This allows straightforward passing of any ground truth and predicted segmentation masks to the *evaluate()* function for computing the metric assessment in a single line of code.

2.3. Package Stability

Our MISeval package utilizes modern DevOps strategies to ensure package stability and functionality during ongoing development [4]. After each update, the source code is automatically built in a reproducible environment, extensively tested via unit testing, released, and, finally, deployed in the scientific community's MIS projects.

Our unit testing considers functionality, edge cases, and exceptions for each metric. For application (functionality and edge cases), multiple dummy dataset types like empty, full, or random segmentation masks as well as single and multi-class masks are tested in all combinations. For exception handling, cases with incorrect parameter usage and non-matching mask shapes are tested.

2.4. Package Availability

The MISeval package is hosted, supported, and version-controlled in the Git repository platform GitHub. This allows the utilization of platform-hosted DevOps workflows and a hub for package documentation, community contributions, bug reporting as well as feature requests. The Git repository is available under the following link: https://github.com/frankkramer-lab/miseval. Furthermore, MISeval is published in the Python Package Index (PyPI), which is the official third-party software repository for Python. Thus, MISeval can be directly installed and immediately used in any Python environment using "*pip install miseval*".

Our code is licensed under the open-source GNU General Public License Version 3 (GPL-3.0 License), which allows free usage and modification for anyone.

3. Results

For qualitative evaluation and functionality demonstration, we setup a deep-learning based MIS pipeline, trained a COVID-19 segmentation model for CT scans, computed predictions, and evaluated model performance using MISeval. The evaluation results are illustrated in Figure 1.The analysis utilized the MIS framework MIScnn [2] with default parameters. As dataset, we used annotated computed tomography scans of COVID-19 positive patients from Ma et al. [5].

For quantitative evaluation, we compared our metric library with other widely used frameworks for machine learning and image analysis. As it can be seen in Table 1,

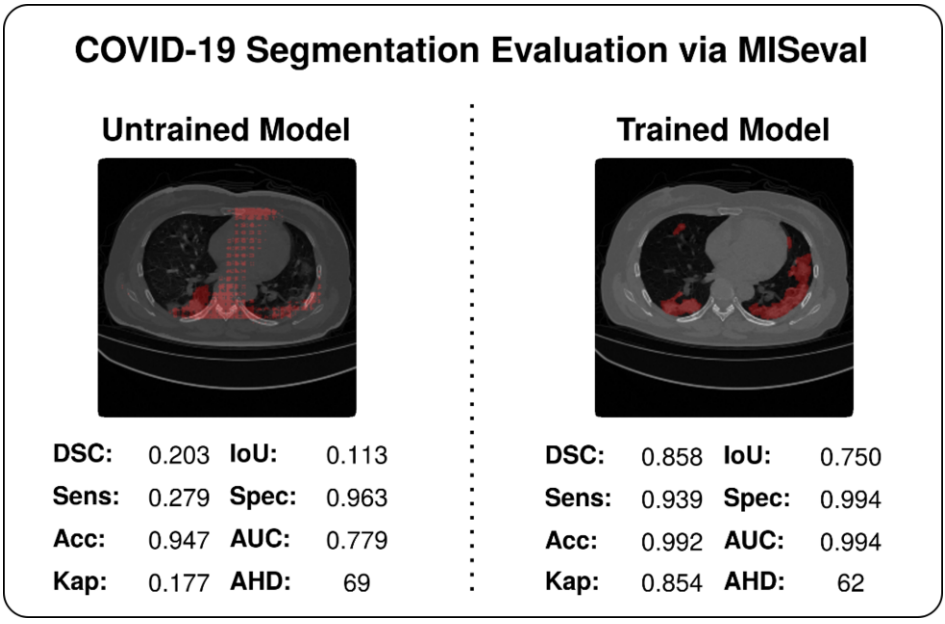


Figure 1. Illustration of various selected metrics from the library of MISeval to evaluate model performance on the use case COVID-19 infected region segmentation. The figure compares an untrained model (after 1 epoch during training) to a fully trained model (after 163 epochs) and shows computed tomography scans for each model with predicted infected regions (red).

MISeval provides currently 28 metrics, which is the highest number of segmentation metrics compared to other analyzed frameworks: scikit-learn [6] with 18, EvaluateSegmentation from VISCERAL [7] with 13, PyMIA [8] with 12, Tensorflow [9] with 16 and TorchMetrics [10] with 12.

Table 1. Overview and comparison of currently implemented metrics in MISeval. For in-detail formula description and theory of the majority of presented metrics, we refer to the review from Taha et al. [4].

Group	Metric	scikit-learn	VISCERAL	PyMIA	Tensorflow	TorchMetrics	MISeval
Spatial Overlap	Dice Similarity Coefficient / F1-score	X	X	X	X	X	X
	Intersection-Over-Union / Jaccard Index	X	X	X	X	X	X
	Sensitivity / Recall	X	X	X	X	X	X
	Specificity		X	X	X	X	X
	Precision	X	X	X	X	X	X
Spatial Distance	(Average) Hausdorff		X	X			X
	Bhattacharyya						X
	Canberra						X
	Chebyshev						X
	Chi Square	X					X
	Cosine	X			X		X
	Euclidean	X					X
	Manhattan	X			X		X
	Hamming	X			X	X	X
	Mahanabolis		X				
	Minkowski						X
	MAE / MSE	X		X	X		X
	Pearson						X
Correlation	Interclass Correlation		X	X			
	Matthews Correlation	X			X	X	X
Divergence	Jensen-Shannon						X
	Kullback-Leibler				X	X	X
	Cross-Entropy	X			X		X
	Hinge	X			X	X	X
Probabilistic or Pairing	AUC	X	X	X	X	X	X
	Cohen Kappa	X	X	X	X	X	X
	Accuracy / Rand Index	X	X	X	X	X	X
	Balanced Accuracy	X					X
	Adjusted Rand Index	X	X	X			X
Volume	Volumetric Similarity		X				X

4. Discussion

Our proposed package MISeval allows a universal, reproducible, and standardized application of various metrics for MIS evaluation, which hopefully reduces the risk of statistical bias in studies through incorrect custom implementations. By following the state-of-the-art package stability and availability strategies, MISeval has the potential to be integrated into any future scientific performance analysis due to package stability, easy accessibility, and further contribution possibilities.

Our road map and future direction for MISeval is to ensure ongoing support, the further extension of our metric library, and providing guidelines on correct metric usage as well as evaluation. Furthermore, we plan to propose a new metric similar to the Dice Similarity Coefficient for handling the current issue of evaluating non-present classes in ground truth annotations like in control samples.

5. Conclusions

In this work, we proposed our open-source Python package MISeval: a metric library for medical image segmentation evaluation. The library contains various popular metrics which can be easily used and integrated into any performance assessment for image segmentation models. MISeval can be directly installed as a Python library from PyPI (miseval) and is available in GitHub: *https://github.com/frankkramer-lab/miseval*.

References

[1] Litjens G, Kooi T, Bejnordi BE, Arindra A, Setio A, Ciompi F, et al. A survey on deep learning in medical image analysis. 2017;42(December 2017):60–88.

[2] Müller D, Kramer F. MIScnn : a framework for medical image segmentation with convolutional neural networks and deep learning. BMC Med Imaging. 2021 Jan 21;21:12.

[3] Zhang Y, Mehta S, Caspi A. Rethinking Semantic Segmentation Evaluation for Explainability and Model Selection.

[4] Dyck A, Penners R, Lichter H. Towards definitions for release engineering and DevOps. In: Proceedings - 3rd International Workshop on Release Engineering, RELENG 2015. Institute of Electrical and Electronics Engineers Inc.; 2015. p. 3.

[5] Ma J, Ge C, Wang Y, An X, Gao J, Yu Z, et al. COVID-19 CT Lung and Infection Segmentation Dataset [Internet]. Zenodo; 2020. Available from: https://doi.org/10.5281/zenodo.3757476

[6] Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine Learning in {P}ython. J Mach Learn Res. 2011;12:2825–30.

[7] Taha AA, Hanbury A. Metrics for evaluating 3D medical image segmentation: Analysis, selection, and tool. BMC Med Imaging [Internet]. 2015 Aug 12 [cited 2021 May 14];15(1):29. Available from: http://bmcmedimaging.biomedcentral.com/articles/10.1186/s12880-015-0068-x

[8] Jungo A, Scheidegger O, Reyes M, Balsiger F. pymia: A Python package for data handling and evaluation in deep learning-based medical image analysis. Comput Methods Programs Biomed. 2021 Jan 1;198:105796.

[9] Abadi M, Agarwal A, Barham P, Brevdo E, Chen Z, Citro C, et al. TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems [Internet]. 2015. Available from: https://www.tensorflow.org/

[10] Detlefsen N, Borovec J, Schock J, Jha A, Koker T, Di Liello L, et al. TorchMetrics - Measuring Reproducibility in PyTorch. J Open Source Softw [Internet]. 2022 Feb 11 [cited 2022 Mar 19];7(70):4101. Available from: http://arxiv.org/abs/1907.11692

Challenges of Trustable AI and Added-Value on Health
B. Séroussi et al. (Eds.)

doi:10.3233/SHTI220392

When Context Matters for Credible Measurement of Drug-Drug Interactions Based on Real-World Data

Romain LELONG[a,b], Badisse DAHAMNA[a,b], Hélène BERTHELOT[a], Willy DUVILLE[a], Catherine LETORD[a,b], Julien GROSJEAN[a,b] and Catherine DUCLOS[a,1]

[a] *Université Sorbonne Paris Nord, Laboratoire d'Informatique Médicale et d'Ingénierie des connaissances en e-Santé, LIMICS, Sorbonne Université, INSERM, F-93000, Bobigny, France*

[b] *Department of Digital Health, Rouen University Hospital, Rouen, France*

Abstract. The frequency of potential drug-drug interactions (DDI) in published studies on real world data considerably varies due to the methodological framework. Contextualization of DDI has a proven effect in limiting false positives. In this paper, we experimented with the application of various DDIs contexts elements to see their impact on the frequency of potential DDIs measured on the same set of prescription data collected in EDSaN, the clinical data warehouse of Rouen University Hospital. Depending on the context applied, the frequency of daily prescriptions with potential DDI ranged from 0.89% to 3.90%. Substance-level analysis accounted for 48% of false positives because it did not account for some drug-related attributes. Consideration of the patient's context could eliminate up to an additional 29% of false positives.

Keywords. drug-drug interaction, prescription, data warehouse, methodology

1. Introduction

Observational studies are becoming more feasible with the computerization of health data. Many studies concerning the SARS-COV2 have used electronic medical records that have improved our knowledge on this new disease. However, the credibility of some of these studies has been questioned as some have been retracted after being published in major journals (1).

Numerous studies have reported the frequency of drug-drug interactions (DDI) among patients, with surprisingly wide-ranging values, from a few tenths to a few tens (2–4). This discrepancy has a multifactorial explanation: origin of data (prescription, reimbursement), perimeter of both data and subjects (inpatient, outpatient, admissions, elderly), definition of the DDIs (referential, DDI checkers, severity), definition of exposure (a posteriori reconstitution of the daily prescription, definition of drug co-occurrence), drug-level used to perform the analysis (clinical drug, active ingredient, ATC class).

[1]Corresponding Author, Catherine Duclos, LIMICS, Université Sorbonne Paris Nord, 74 rue Marcel Cachin, Bobigny, France; E-mail: catherine.duclos@inserm.fr

For a good detection of potential DDIs, a more contextualized definition of DDIs is also important and limits false positives (5). The context can be divided into drug-related factors (e.g. dosage, route of administration, formulation) or patient-related factors (e.g. comorbidities, elimination phenotypes) (6) but experiments of contextualized detection have been only performed on a few DDIs.

In this paper, we use a unique prescription dataset from the Rouen University Hospital data warehouse (7) to test the impact of different contextual items on the DDI frequency measurement.

2. Material and Methods

2.1. Identification of the manufactured pharmaceutical products potentially eligible for a DDI.

An initial list of Proprietary Medicinal Products (PMP) marketed in 2017 was compiled with the available data in the Répertoire des Spécialités Pharmaceutiques (http://agence-prd.ansm.sante.fr/php/ecodex/). These PMP were identified by their specific French code name "Code Identifiant de Spécialité" (CIS, Proprietary Medicinal Product identifier). The French Thesaurus of DDIs (8) was the reference material for contra-indicated DDI. These DDIs were manually analyzed by two pharmacists to identify the context attributes that might narrow the scope of a DDI to a subset of either substances, or PMPs or patients. The PMPs description was enriched with the attributes of interest through the information contained in their Summary of Product Characteristics. The initial list of PMPs was then reduced to those containing the substances referred in DDIs (without further filtering). For each DDI, the list of eligible PMPs was filtered by the context attributes applicable for this DDI. Finally, a combinatorial calculation was performed to create all PMPs pairs whose substance component belongs to either the object or the precipitant of the DDI. This list constituted the List 1 of all theoretical PMPs pairs potentially involved in a DDI, without any filter. This first list was then reduced according to the different applicable filters and an additional filter of no interaction between the substance and itself was applied on the PMPs pairs. This final list was considered the gold standard list of PMPs pairs.

2.2. Data extraction process

Medications electronically prescribed for patients hospitalized in 2017 in Rouen University Hospital were extracted from EDSaN data warehouse in accordance with current regulations on health data privacy (anonymous, partial, aggregated data with a minimum threshold of ten). For reasons of parsimony, only daily prescription lines containing a PMP likely to be involved in a DDI were included in the study. As drugs were identified by their French code name "Unité Commune de Dispensation" (UCD, common dispensation unit), a mapping between the CIS and the UCD was performed using the multilingual terminology server HeTOP (9) which integrated the UCD repository provide by the French Agency of Digital Health (https://esanté.gouv.fr/) (10). The metadata extracted were: the anonymous patient identifier, the anonymous hospital stay identifier, the anonymous daily prescription identifier and the drugs identified by

their UCD. A DDI was only considered if the two drugs involved coexisted in the same daily prescription (DP).

2.3. Analysis

The number of DPs was calculated on the whole 2017 prescriptions and was used as the denominator of the frequency indicators. For each DP, all drugs pairs were generated and compared to the list 1 of PMPs pairs. Only the DPs containing at least one PMPs pair involved in a DDI were retained and their number was calculated. The number of DPs with DDIs was then recalculated by applying the filters corresponding to each studied context to the prescription dataset. These numbers are the numerators of the frequency indicators. The percentage of false positives was defined as the number of falsely detected PMPs pairs out of the whole detected PMPs pairs.

3. Results

A total of 256 contraindicated DDIs were studied. A DDI could occur between two classes of substances (e.g. *irreversible monoamine oxidase inhibitors and monoamine oxidase-metabolized triptans*), a class and a substance (e.g. *statins and fusidic acid*), a class with itself (e.g. *fibrates and other fibrates*), or between two substances (*e.g. gemfibrozil and dasabubir*). Hierarchical relationships existed between some substances and classes (e.g. *rasagiline and selegiline are members of MAOI-B class*). Substances were either active or inactive components of PMPs (*e.g. sorbitol, alcohol*).

Among the 11,221 PMPs marketed in 2017, 5,032 PMPs had potential involvement in a DDI; 194,866 theoretical PMPs pairs involved in a DDI were formed by combinatorial association of PMPs belonging to either the object or the precipitant of the DDI (list 1 of PMPs pairs). For example the 479 PMPs related to *statins* could theoretically interact with the 20 PMPs related to *fucidic acid* and led to 7,664 PMP pairs. Of these 194,866 PMPs pairs, 92,691 could be excluded from the list 1 because they did not fit the contextual description expected in the DDI definition such as (i) ineligible dosage (*e.g. acetylsalicylate dosage < 500mg*) (ii) ineligible route (*e.g. nasal*) (iii) bioavailability issue (*e.g. drug without systemic effect*) (iv) substance restriction (*e.g. warfarin is the only substance concerned by a DDI on anticoagulants*) (v) or ineligible indication (*e.g. beta blockers that don't have heart failure as an indication*). In the preceding example, only 4 PMPs related to *fucidic acid* had systemic effect, so only 1,916 PMP pairs were retained for the gold standard for the DDI between *statins* and *fucidic acid*. Of the remaining 102,175 PMPs pairs (gold standard), 14,140 could be further filtered using patient data such as history (e.g. *history of gastrointestinal ulcer*), or biology (e.g. *hypokalemia*).

Of the 5,032 expected CIS, 4,840 were mapped to an UCD using HeTOP. The lack of mapping was due to a lack of coverage of the repository provided by the French regulatory authority.

The dataset extracted from EDSaN contained 916,584 DPs for 2017. Within these prescriptions, only 1,963 of the UCDs were of interest and only 333 of these were co-occurring as a pair of DDI. In the end, only 990 different PMPs pairs from the list 1 were identified.

Depending on the different context filters to be applied, the frequency of DPs with a DDI varied as presented in Table 1. Without any filter, the percentage of DPs with a

potential DDI was 3.9%. When drug context restrictions were applied, the percentage decreased to 2.1%, and with the assumption of a required but unmet patient context, the percentage dropped to 0.89%.

Substance-level analysis accounted for 48% of falsely detected PMP pairs that could be corrected by filtering by drug attributes. Consideration of the patient's context could eliminate up to an additional 29% of false positives. A DP containing initially a potential DDI was detected for 3,068 patients and only for 60% of them when the drug related filters were applied.

Among the 916,584 DPs, 35,798 contained at least a potential DDI pair from the list 1. After applying the successive filters, up to 77% of them could be eliminated.

Table 1. Evolution in DDI frequency taking contextual elements into account and percentage of DP considered falsely detected

Applied Filter	Percentage of DP with a DDI	% of DP eliminated because falsely detected
Without any filter	3.90%	-
Suppress PMPS with a substance interacting with itself	3.26%	16%
Restrict to substances subsets of certain DDIs	2.86%	10%
Restrict to eligible dose forms	2.67%	5%
Restrict to eligible indications	2.15%	14%
Restrict to eligible dosage	2.10%	1%
Restrict to systemic bioavailability	2.10%	0%
Apply patient related filters	0.89%	up to 31%

4. Discussion and Conclusion

In the reported experience, we showed that, with the same dataset, the measurement of the frequency of potential DDIs can vary with a fourfold, depending on the application of contexts items described in the DDI thesaurus.

The right definition of what drugs should be included in a DDI is a first statement. These drugs cannot be solely defined by their substance or ATC class because it is sometimes too broad when forms or dosage matter. This would lead to an overestimation of the frequency of DDIs. The generic drug definition (substance + dosage +form) is sometimes not suitable when excipient or indication matter. This would lead to an underestimation of the frequency of DDIs. PMP seems to be the right level of analysis, but its scope is national which makes international comparisons difficult. The second statement is the importance of patient context in the analysis of DDIs. As we did not extract these patients content, we made some hypothesis about their existence. First we assumed that PMPs were used according to their regulatory authorization (for the indication) and were administrated. This may also have overestimated the frequency of potential DDIs. Second we made the hypothesis they were present. But in real world, we would have to deal with what explicitly exists or not and what is not explicitly stated. Assuming that something that is not explicitly stated is exists would lead to either an overestimation or an underestimation of the frequency of DDIs. Our experiment has some limitations. Prescription data were limited to the drugs prescribed and no other attributes were available (such as the administered dose, the method of administration...). It does not reflect the real exposure

of the patient and may result in an overestimate. Only drugs prescribed on the same day were considered, which may have underestimated the frequency of potential DDIs, because some drugs such as monoamine oxidase inhibitors can interact for up to 15 days after discontinuation. After filters were applied, the remaining DDIs could be considered as true positives, but the link to adverse events should be researched to be affirmative. To avoid contesting the national referential of DDIs (11), information on evidence and incidence should be provided (12) .

This study confirms the importance of context in limiting a large proportion of false positives in DDI detection (5). This gives useful guidance to conduct studies that attempt to determine the frequency of DDI on real-world data or for critical appraisal of such studies. This methodology could be reused to produce indicators on prescription containing contra-indicated DDI and to help pharmacists to improve their quality process.

Acknowledgments

This study was partially granted by (a) QLAP (Quantification des perofmances des logiciels d'aide à la prescription pour la détection des intéractions médicamentueses) ANSM – Appel à projet 2016 – Projet 7, (b) ReMIAMes (Ressources et méthodes libres pour la détection des interactions médicamenteuses dans les entrepôts de données de santé) ANR-19-CE19-0025-01

References

[1] Kohane IS, Aronow BJ, Avillach P, Beaulieu-Jones BK, Bellazzi R, Bradford RL, et al. What Every Reader Should Know About Studies Using Electronic Health Record Data but May Be Afraid to Ask. J Med Internet Res. 2021;23(2):e22219.

[2] Fung KW, et al. Comparison of three commercial knowledge bases for detection of drug-drug interactions in clinical decision support. J Am Med Inform Assoc JAMIA. 2017;24(4):806-12.

[3] Gagne JJ, Maio V, Rabinowitz C. Prevalence and predictors of potential drug-drug interactions in Regione Emilia-Romagna, Italy. J Clin Pharm Ther. 2008;33(2):141-51.

[4] Neto PRO, Nobili A, Marusic S, Pilger D, Guidoni CM, Baldoni A de O, et al. Prevalence and Predictors of Potential Drug-Drug Interactions in the Elderly: A Cross-Sectional Study in the Brazilian Primary Public Health System. J Pharm Pharm Sci. 2012;15(2):344-54.

[5] Chou E, Boyce RD, Balkan B, Subbian V, Romero A, Hansten PD, et al. Designing and evaluating contextualized drug–drug interaction algorithms. JAMIA Open. 2021 ;4(1):ooab023

[6] Horn J, Ueng S. The Effect of Patient-Specific Drug-Drug Interaction Alerting on the Frequency of Alerts: A Pilot Study. Ann Pharmacother. 2019;53(11):1087-92.

[7] Pressat-Laffouilhère T, Balayé P, Dahamna B, Lelong R, Billey K, Darmoni SJ, et al. Evaluation of Doc'EDS: A French Semantic Search Tool to Query Health Documents from A Clinical Data Warehouse. 2020. Disponible sur: https://hal.archives-ouvertes.fr/hal-03066666

[8] Agence nationale de sécurité du médicament et des pro-, duits de santé. Thésaurus des interactions médicamenteuses [Internet]. Disponible sur: https://ansm.sante.fr/documents/reference/thesaurus-des-interactions-medicamenteuses-1

[9] Grosjean J, Merabti T, Griffon N, Dahamna B, Soualmia LF, Darmoni S. Multi-terminology cross-lingual model to create the Health Terminology/Ontology Portal. TIA. 2011; 119-122

[10] Grosjean J, Letord C, Charlet J, Aimé X, Danès L, Rio J, et al. Un modèle sémantique d'identification du médicament en France. Atelier IA et Santé, 2019.

[11] Bensalah N, et al. Conduite à tenir en cas d'association d'un alpha-bloquant et d'un inhibiteur puissant du cytochrome 3A4. Prog En Urol. 2017; 27(5):275-82.

[12] Brochhausen M, et al. Towards a foundational representation of potential drug-drug interaction knowledge. CEUR Workshop Proc. 2014;1309:16-31.

Challenges of Trustable AI and Added-Value on Health
B. Séroussi et al. (Eds.)

doi:10.3233/SHTI220393

A Lightweight and Interpretable Model to Classify Bundle Branch Blocks from ECG Signals

Hugues TURBÉ[a,1], Mina BJELOGRLIC[a], Mehdi NAMDAR[b], Christophe GAUDET-BLAVIGNAC[a], Jamil ZAGHIR[a], Jean-Philippe GOLDMAN[a], Belinda LOKAJ[a] and Christian LOVIS [a]
[a] *Division of medical information sciences, University hospitals of Geneva and Department of radiology and medical informatics, University of Geneva, Switzerland*
[b] *Cardiology Division, University Hospitals of Geneva, Switzerland*

Abstract. Automatic classification of ECG signals has been a longtime research area with large progress having been made recently. However these advances have been achieved with increasingly complex models at the expense of model's interpretability. In this research, a new model based on multivariate autoregressive model (MAR) coefficients combined with a tree-based model to classify bundle branch blocks is proposed. The advantage of the presented approach is to build a lightweight model which combined with post-hoc interpretability can bring new insights into important cross-lead dependencies which are indicative of the diseases of interest.

Keywords. ECG automatic classification, Interpretability, Lightweight Model

1. Introduction

An electrocardiogram is a very common clinical exam, low-cost, non-invasive and allow to diagnose a range of cardiovascular diseases. Electrocardiograms measure the electric activity of the heart and aim to build a map of the heart in three orthogonal directions. Standard ECGs produce a signal with 12 derivations measured over approximately 10 seconds with a standard sampling frequency of 500Hz resulting in series with 5000 timesteps. ECGs are often analysed by non-experts staff and reports suggest that they are open to misdiagnosis [1]. This data abundance and a need for accurate diagnostic created a longtime interest for the automatic classification of diseases from electrocardiogram signals. Classification of time-series has often been seen as a difficult task for neural networks which struggle to learn long-term dependencies often encountered in ECG data. Recently, numerous models, mainly based on deep-neural networks, have achieved very high classification accuracy for a range of cardiovascular diseases [2,3].

One drawback of these approaches based on deep neural network is the lack of interpretability of the developed models. In this work, interpretability refers to building an understanding of which features were used and their relative importance for the model to classify a given ECG. Indeed with ever more complex models, it can be very difficult

[1] Corresponding Author: Hugues Turbé; E-mail: hugues.turbe@unige.

to understand how a model came to a given conclusion on a specific sample. Different methods have been developed to address this issue and provide what has been termed post-hoc explainability [4]. Post-hoc explainability is a key requirement for clinicians to adopt these technologies [5], but might also become a legal requirement following a recent recommendation from the European Commission [6]. Recent works looked at providing interpretability in the context of ECG classification using deep neural networks (DNN). The current approach consist in first training a DNN and then applying post-hoc explainability, principally LIME [7], to highlight the most important part of the signal used by the network to provide the classification of a given sample. The LIME interpretability method is widely used as it is model-agnostic, meaning it can be applied on any architecture. LIME and other interpretability methods were initially developed for natural language processing (NLP) or image classification task. Given this background, one current limitation is that they fail to highlight important dependencies captured by the network both across time steps and different leads of the ECG. With this issue in mind, the presented method aims to address the following issues:

- The computing intensive nature of deep neural network required to learn the complex dependencies observed in ECG signals.
- The lack of interpretability method to highlight dependencies across leads captured by the models.

As part of this research, it was decided to build models to classify the presence of left and right bundle branch blocks respectively LBBB and RBBB. Bundle branch blocks refer to delay or blockage of the electrical signal responsible for making the heart beat [8]. Bundle branch blocks are associated with higher risk of serious cardiovascular complications and mortality in specific conditions [9,10]. While automatic classification of bundle branch blocks is well studied [11], the following work aims to present a much smaller model, which does not rely on CNN or LSTM layers, with improved interpretability of the classification models.

2. Methods

The CPSC subset from the large *PhysioNet: Classification of 12-lead ECGs* dataset [12] was used. This dataset includes a total of 6850 ECGs, with a minimal length of 10 seconds, annotated for 9 cardiovascular diseases. Out of these ECGs, 232 were annotated with a LBBB diagnostic, 1854 with a RBBB and 916 with a normal rhythm.

The model presented in this work is based on using a multivariate autoregressive model (MAR) as a preprocessing step to capture inter-lead dependencies. MAR model aims to explain current values of a multivariate time series as a linear combination of the past values. A MAR model of order p will use the p previous time step of the serie of interest to predict the following. For a time series $\boldsymbol{X} \in \mathbb{R}^{M \times T}$ where M is the number of feature of the time series, in our case the number of derivation obtained with the ECG signals, and T the number of time steps in the series, we can write our MAR model as follows:

$$\boldsymbol{X_t} = \sum_{l=1}^{p} \boldsymbol{A_i} \cdot \boldsymbol{X_{t-i}} + \epsilon_t \quad (1)$$

Where $\boldsymbol{A_i} \in \mathbb{R}^{M\times M}$ is the matrix of coefficient and ϵ_t the additive Gaussian noise with zero mean and covariance R. For the rest of the analysis, a MAR model of order one is used, so that $\boldsymbol{A} = \boldsymbol{A_i}$. The matrix $\boldsymbol{A}$ is central to the presented analysis as it encodes the conditional dependencies across the leads of the ECG signals [13] as well as reducing the dimension from the domain $\mathbb{R}^{5000\times12}$ to a smaller $\mathbb{R}^{12\times12}$ domain. This dimensionality reduction allows to greatly reduce the computing power required to train a classification model. The rest of the analysis relies on the assumption that the changes induced by a left or right bundle branch will affect the dependency across the leads and hence can be used as a marker to classify these two diseases. MAR models have been widely adopted to infer organisation model of the human brain using functional magnetic resonance imaging (fMRI) time series [14]. Surprisingly these models have not yet been used, to the authors' knowledge, on ECG data. As a prerequisite to compute the MAR coefficients, the stationarity of the time series was tested using the Augmented Dickey–Fuller test. In addition, all samples were normalised by removing their mean and scaling them to unit variance.

A first statistical analysis of the significance of the MAR model's coefficients with regards to the diseases of interest was performed using a permutation test based on the means of the groups [15].

For each of the disease of interest, the MAR coefficients were used as input to train a tree-based model, more specifically a Light Gradient Boosting Model (LightGBM) [16]. LightGBM is based on gradient boosted tree model and has gained attraction for its state-of-the-art performance. For each of the two trained models the target for each sample was whether the latter was annotated with the disease of interest or as a sample with a sinus rhythm. Once these models are trained, *Tree SHAP* [17] is used to provide post-hoc interpretability under the form of Shapley values. Shapley values aim to explain which coefficients of the MAR model were used by the trained tree model as indicative of the diseases of interest.[2]

3. Results

Initial permutation tests showed that out of the 144 (12×12) MAR coefficients, 66 for LBBB and 84 for RBBB had a significant distribution difference with regards to samples with a sinus rhythm $p \leq 0.05$. Hyperparameter optimisation was performed for the two classification tasks with the optimal hyperparameters for each model along resulting classification metrics presented respectively in Table 1 and Table 2.

Table 1. Model's Hyperparameters

	LBBB	RBBB
Learning rate	0.12	0.11
Nb. leaves	1420	1220
Max depth	4	6

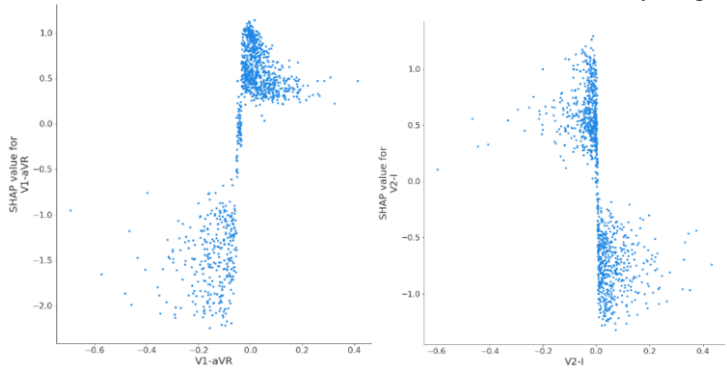
Table 2. Classification Metrics

	LBBB	RBBB
Accuracy	0.97	0.84
Precision	0.95	0.85
Recall	0.88	0.91
F1	0.91	0.88

Shapley values were calculated using TreeShap [17] for the two models. Shapley values explain how each feature across the dataset influenced the prediction of the model of

[2] Full implementation of the discussed method is presented at https://github.com/hturbe/ECG_MAR_Model

interest. In Figure 1, the influence of two important MAR coefficients on the classification of LBBB are shown. This figure shows the Shapley value attributed to a specific feature across the entire dataset and allows to gain insights into how specific values of the feature influence the model's predictions. The magnitude of the Shapley value shown on the y-axis reflects the importance of this feature for the model to classify the sample. In addition, a positive Shapley value indicates that this feature played in favour of the sample being classified as having the disease of interest while a negative Shapley value indicates that the feature played against the sample being predicted with the disease. Figure 1a shows that a positive dependence between lead V1 and aVR is associated by the network as an indication of a LBBB while the opposite is true for the dependence between leads V2 and I as shown by Figure 1b. Regarding RBBB, the dependence between leads V1 and V2 was found to be a key diagnostic criterion.



(A) (B)

Figure 1. Shapley values computed on the model trained to predict LBBB. Shapley values are presented for the MAR coefficient encoding the dependence between leads V1-aVR (a) and leads V2-I (b).

4. Conclusions

The presented approach is based on encoding the temporal dependencies across the ECG leads within the MAR model. This initial step allows to reduce each ECG into a matrix of 144 coefficients. The coefficients are then used as inputs for tree-based models with smaller computing requirements than traditional approaches based on neural networks architectures. Two models are trained achieving an accuracy of respectively 97% for the classification of LBBB and 84% for RBBB. Of interest, the accuracy for the classification of LBBB is achieved with a dataset including only 232 samples annotated with the disease. One current drawback of the method is that it discards "intra-lead" feature, such as QRS length and others. An avenue to fix this issue would be to augment the MAR coefficients with intra-signal measurements.

Shapley values provide interesting insights into how a change in the dependence between a pair of leads are indicative of a specific disease. The research showed how change in dependencies between lead V1 and aVR and leads V2 and I were discriminative to predict a LBBB. In addition, while some research have focused on developing interpretability methods for ECG classification, they show time steps in given leads which were important for the classification. This information is interesting but does not reflect how dependence across leads, which might be captured by the developed models, play a role in the final prediction.

Finally, the models presented in this article should be trained and tested on a more extensive dataset. Indeed, currently the classification is limited between sinus rhythm and the two diseases which were analysed. However this classification does not reflect a clinical setting where patient might also exhibit different cardiovascular diseases.

References

[1] Awan M, Omar M, Omar BA, Malozzi C, Awan GM. Abstract 244: ECG Recording Errors Prevalence in a Hospital Setting, *Circulation: Cardiovascular Quality and Outcomes*. **11** (2018) A244–A244. doi:10.1161/circoutcomes.11.suppl_1.244.

[2] Ribeiro AH, Ribeiro MH, Paixão GMM, Oliveira DM, Gomes PR, Canazart JA, et al. Automatic diagnosis of the 12-lead ECG using a deep neural network, *Nat Commun*. **11** (2020) 1760. doi:10/drkd.

[3] Liang Y, Yin S, Tang Q, Zheng Z, Elgendi M, Chen Z. Deep Learning Algorithm Classifies Heartbeat Events Based on Electrocardiogram Signals, *Front. Physiol.* **11** (2020) 569050. doi:10/gm3srb.

[4] Lipton ZC.The Mythos of Model Interpretability: In machine learning, the concept of interpretability is both important and slippery., *Queue*. **16** (2018) 31–57. doi:10/gkbfrd.

[5] Shad R, Cunningham JP, Ashley EA, Langlotz CP, Hiesinger W. Designing clinically translatable artificial intelligence systems for high-dimensional medical imaging, *Nat Mach Intell*. **3** (2021) 929–935. doi:10/gnn5hz.

[6] European Commission, Directorate-General for Communications Networks, Content and Technology, Proposal for a regulation of the European parliament and of the council laying down harmonised rules on artificial intelligence (artificial intelligence act) and amending certain union legislative acts COM/2021/206 final, (2021).

[7] Ribeiro MT, Singh S, Guestrin C. "Why Should I Trust You?": Explaining the Predictions of Any Classifier, *ArXiv:1602.04938 [Cs, Stat]*. (2016). http://arxiv.org/abs/1602.04938 (accessed March 2, 2021).

[8] McAnulty JH, Rahimtoola SH. Bundle branch block, *Progress in Cardiovascular Diseases*. **26** (1984) 333–354. doi:10.1016/0033-0620(84)90009-4.

[9] Tan NY, Witt CM, Oh JK, Cha YM. Left Bundle Branch Block: Current and Future Perspectives, *Circ: Arrhythmia and Electrophysiology*. **13** (2020). doi:10.1161/CIRCEP.119.008239.

[10] Bussink BE, Holst AG, Jespersen L, Deckers JW, Jensen GB, Prescott E. Right bundle branch block: prevalence, risk factors, and outcome in the general population: results from the Copenhagen City Heart Study, *European Heart Journal*. **34** (2013) 138–146. doi:10.1093/eurheartj/ehs291.

[11] Oh SL, Tan EYKNgRS, Acharya UR. Automated diagnosis of arrhythmia using combination of CNN and LSTM techniques with variable length heart beats, *Computers in Biology and Medicine*. **102** (2018) 278–287. doi:10/gfnqsp.

[12] Perez Alday EA, Gu A, Shah A, Robichaux C, Wong AKI, Liu C, et al. Classification of 12-lead ECGs: the PhysioNet/Computing in Cardiology Challenge 2020, Cardiovascular Medicine, 2020. doi:10.1101/2020.08.11.20172601.

[13] Lütkepohl H, New introduction to multiple time series analysis, Springer Science & Business Media, 2005.

[14] Rogers BP, Katwal SB, Morgan VL, Asplund CL, Gore JC. Functional MRI and multivariate autoregressive models, *Magnetic Resonance Imaging*. **28** (2010) 1058–1065.

[15] Wilcox RR, 8 - COMPARING TWO INDEPENDENT GROUPS, in: R.R. Wilcox (Ed.), Applying Contemporary Statistical Techniques, Academic Press, Burlington, 2003: pp. 237–284. doi:10.1016/B978-012751541-0/50029-8.

[16] Ke G, Meng Q, Finley T, Wang T, Chen W, Ma W, et al. LightGBM: A Highly Efficient Gradient Boosting Decision Tree, in: I. Guyon, U.V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Eds.), Advances in Neural Information Processing Systems, Curran Associates, Inc., 2017. https://proceedings.neurips.cc/paper/2017/file/6449f44a102fde848669bdd9eb6b76fa-Paper.pdf.

[17] Lundberg SM, Erion G, Chen H, DeGrave A, Prutkin JM, Nair B, et al. From local explanations to global understanding with explainable AI for trees, *Nature Machine Intelligence*. **2** (2020) 56–67.

Challenges of Trustable AI and Added-Value on Health
B. Séroussi et al. (Eds.)

doi:10.3233/SHTI220394

Analysis of Stroke Assistance in Covid-19 Pandemic by Process Mining Techniques

Gabrielle dos Santos LEANDRO[a,b,1], Daniella Yuri MIURA[c], Juliana SAFANELLI[b], Rafaela Mantoan BORGES[c] and Cláudia MORO[a]
[a] *Pontifícia Universidade Católica do Paraná, Curitiba, Brazil*
[b] *Joinville Stroke Registry, Joinville, Brazil*
[c] *Lehigh University, Bethlehem, EUA*

Abstract. Medical assistance to stroke patients must start as early as possible; however, several changes have impacted healthcare services during the Covid-19 pandemic. This research aimed to identify the stroke onset-to-door time during the Covid-19 pandemic considering the different paths a patient can take until receiving specialized care. It is a retrospective study based on process mining (PM) techniques applied to 221 electronic healthcare records of stroke patients during the pandemic. The results are two process models representing the patient's path and performance, from the onset of the first symptoms to admission to specialized care. PM techniques have discovered the patient journey in providing fast stroke assistance.

Keywords. Process Mining, Stroke, Patient Journey

1. Introduction

Stroke is a non-communicable chronic disease that represents the second leading cause of mortality worldwide. Besides, it is considered the second leading cause of disability-adjusted life years loss globally and the main cause of hospitalization in the Brazilian Public Unified Health System[1]. In stroke treatment, a therapeutical window represents the best moment to intervene in the pathological process of cerebral ischemia and minimize damage to the central nervous system [2]. However, this window often lasts only a few hours, which implicates the need for fast medical assistance for people who experience an acute stroke. Furthermore, one of the intervals in this therapeutic window is the onset-to-door time, representing the interval between the onset of stroke symptoms and the patient's hospitalization [3].

Nevertheless, due to the Covid-19 pandemic, drastic and rapid changes have affected the daily operation of healthcare services, such as the suspension of programmed clinical activities during the lockdown period, significantly diminishing the search for emergency medical services caused by fear of contracting the SARS-CoV-2 coronavirus, which resulted in the postponement of appropriate treatment for cerebrovascular diseases. Besides, in this period, assistance to stroke patients suffered negative impacts worldwide, reflecting a sharp reduction in the number of hospitalizations for acute stroke and,

[1] Corresponding Author at: Graduate Program in Health Technology (PPGTS), Pontifícia Universidade Católica do Paraná (PUCPR), 1155 Imaculada Conceição St., Curitiba, PR, 80215-901, Brazil.
E-mail: gabrielle.leandro@pucpr.edu.br

consequently, in the number of patients receiving treatment within the therapeutic window [4].

Thus, the objective of this study was to identify the onset-to-door time through process mining techniques, considering the several paths within the healthcare network during the Covid-19 pandemic. This analysis is relevant because recognizing signs and symptoms and the early calling of emergency services can change the outcome of stroke patients, reducing possible sequelae. Furthermore, process mining (PM) techniques were used to identify and analyze the paths of stroke assistance and its outcomes. The application of PM is helpful since it allows manipulation of data from healthcare information systems, making it possible to analyze the flows and processes of a service, easing the recognition of eventual flaws, and allowing improved coordination of the patient's care [5,6].

2. Methodology

We performed a retrospective cross-sectional study by analyzing event logs from 221 electronic healthcare records of stroke patients in a public hospital in Joinville, Brazil, from March 17 to August 21, 2020. This initial date represents the start of Santa Catarina's social distancing measures. The Brazilian Ethics and Research Committee approved this study as report number 4.917.962, dated August 19, 2021.

The PM methodology applied to this study is composed of five stages (0 to 4) based on the "Process Mining Manifesto" [7]. Stage 0 includes planning and justification, according to our presentation in the Introduction of this research. Stage 1 embraces the data selection and extraction from the health information system [7]. The dataset consisted of the variables: date and time of the first symptoms, date and time when the patient sought help, where the assistance was requested, mode of transport to a hospital that is a reference for stroke assistance, data and time of hospitalization, and the modified Rankin Scale (mRS) of the patient at discharge. The mRS measures the degree of disability in daily activities. We grouped the mRS score according to the ranges: 0-1 (no disability and no significant disability), 2-3 (slight to moderate disability), 4-5 (moderate to severe and severe disability), and 6 (death) to simplify the presentation of results [8].

In Stage 2, after data extraction, data were pre-processed, excluding missing, incomplete, and undefined values in an Excel spreadsheet. The following variables were clustered: date and time of hospitalization and mRS score after discharge; this was grouped to understand the relationship between a patient's possible paths and outcome. An Excel file was input into the Disco software from Fluxicon to discover the process model, and after the file was imported, the columns were assigned to 'caseID,' 'timestamps,' and 'activity names.' The Disco Miner combined Fuzzy Miner with process metrics and modeling strategies[9]. In Stage 3, other process perspectives can be analyzed, such as date, time, and resources. Disco allows the analysis from an instance to a macro process. We presented only a macro process in this study. Finally, the models constructed in Stage 3 can be used in Stage 4 for making interventions, predictions, and recommendations [7]. In the Results of this paper, we presented Stage 2 and 3 outcomes.

3. Results

Figures 1 and 2 present the two discovered process models. These models represent the patient's path from the onset of symptoms to the moment of his admission to a public-stroke reference institution. In Fig. 1 and 2 legends, the inverted triangle indicates the start of the flowchart, the rectangles represent the events (which vary in color according to the number of patients that transit through them: the "lighter" color indicates fewer patients, and the "reddish" color indicates an increased number of patients transiting through that event), the arrows represent the path followed (which change in their width and color according to the intensity of patients transiting through that path), the numbers inside the rectangles or next to the arrows represent the number of cases, that is, the number of patients that went through that path, respectively. Finally, the square inside the red circle indicates the end of the flowchart.

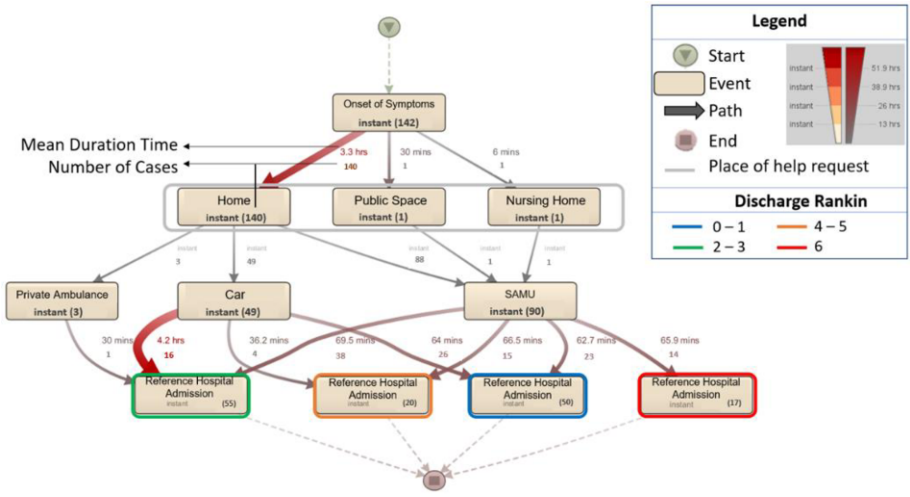


Figure 1. Model 1: Medical assistance requested from home, public space, and nursing home.

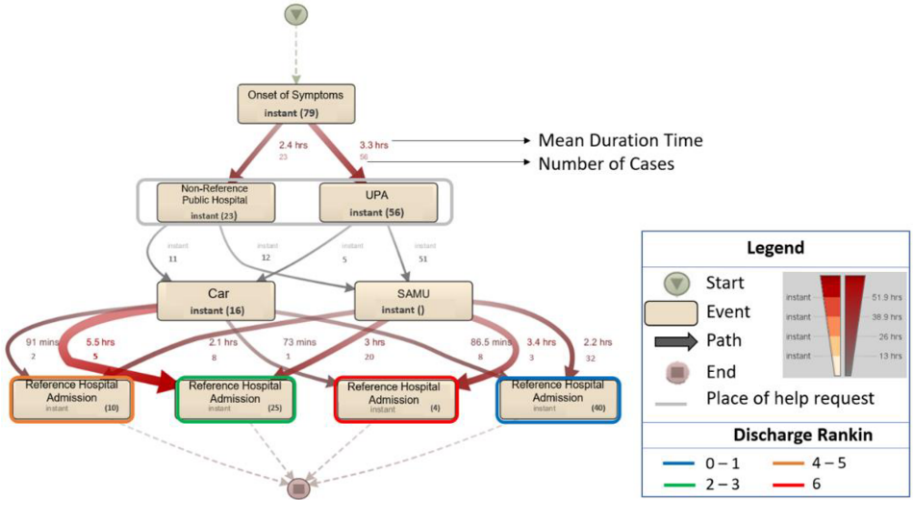


Figure 2. Model 2: The patient went to a Non-Stroke Reference-Public Hospital and UPA (emergency care units).

Fig. 1 presents events where the patient requested assistance at home, in public spaces and in nursing homes; Fig. 2 shows the patient already requested care and went to a Non-Stroke Reference Public Hospital and UPA (UPA are three local emergency care units) after the onset of symptom. In Fig. 1, 142 cases were analyzed, generating 568 events distributed into 11 activities. Some paths were hidden through filters to avoid incomprehensible models as performed in studies [10]. The mean onset-to-door time was 4.7 hours, and the median was 2 hours. Therefore, it is possible to conclude that the mean time between the onset and the assistance requested by patients at home (n = 140) was 3.3 hours, whereas to get to the stroke reference care: the majority used mobile emergency medical services (SAMU) (63%, n = 88), 35% (n = 49) used a private car, and the minority was transported via private ambulance (2%, n = 3). Post-assistance was also analyzed at discharge, 35% (n = 50) the patients evaluated as mRS 0-1, 39% (n = 55) evaluated as mRS 2-3, 14% (n = 20) evaluated as mRS 4-5 and 12% (n = 17) died.

The second model displays the flowchart of patients that, after experiencing the first stroke symptoms, sought help and were referred to public healthcare institutions other than the stroke reference hospital. In this model, 79 cases were analyzed and generated 316 events distributed into nine activities, making it feasible to display all the activities and routes followed by the patients until hospitalization. The mean onset-to-door time was 5.6 hours, and the median was 3.4 hours. Fig. 2 shows that the emergency care unit (UPA) (70%, n = 56) was most commonly sought was by the population among the public non-stroke reference institutions, followed by two other public hospitals (30%, n = 23), where there was 3.3 and 2.4 hours of mean medical assistance time, respectively. In these situations, 9% of the patients in UPA (n = 5) and 48% at the public non-stroke reference hospitals (n = 11) were not referred to the public stroke reference hospital via SAMU or fire-rescue squad but used their private cars.

When referring to the mode of transport compared to the patient's clinical outcome: a Rankin score of 0-1 refers to a mean interval of 2.2 hours (n = 32) via SAMU and 86 minutes (n = 8) via private car. A Rankin score of 2-3 refers to a mean interval of 3 hours (n = 20) via SAMU and 5.5 hours (n = 5) via private car. Rankin score of 4-5 with a mean interval of 2.1 hours (n = 8) via SAMU and 5.5 hours (n = 5) via private car, and Rankin score 6 with a mean interval of 91 minutes (n = 2) via SAMU and 73 minutes via private car.

4. Discussion and Conclusion

PM has made it possible to discover and analyze two process models: the first, the onset-to-door meantime, was 4.7 hours; and the second was 5.6 hours. The Teo et al. (2020) study compared the median stroke onset-to-door time before and during the Covid-19 pandemic, they noted that during the pandemic, the onset-to-door time was ≈1-hour longer than before the pandemic (154 versus 95 minutes, P=0.12), and the proportion of individuals with onset-to-door time within 4.5 hours was significantly lower (55% versus 72%, P=0.024) [11]. Leandro et al. (2022) realized that during the pandemic, the worsening of the patient health status during hospital admission, decreased hospitalization time, increased delay in receiving reperfusion therapies, and preference for the referral hospital over emergency services [6]. Although primarily, the onset-to-door time was maintained within the appropriate interval, it was identified that 36% of the population still sought non-reference-stroke institutions that increased the onset-to-door time, similar to the results found by Nguyen et al. (2021).

Most of the time, in this research, the patient's mode of transport did not influence the onset-to-door time; the patients remained within the therapeutic window of 4.5 hours. However, some medical assistance was not supplied within the therapeutic stroke window, either due to the failure to recognize stroke symptoms or seeking non-specialized care [12]. Therefore, it is necessary to promote campaigns to warn of the stroke signs and symptoms, stress the importance, and guide non-stroke reference hospitals. Even though our study was limited due to the small number of subjects, PM has proven to be a valuable tool for assisting managers in quicker identification of flows to reduce decision-making time.

References

[1] BRASIL. Ministério da Saúde. Health Brazil 2018: An analysis of the health situation and of chronic diseases and conditions: challenges and perspectives [Internet]. Brazil; 2019 [cited 2021 Jan 14]. Available from: https://bvsms.saude.gov.br/bvs/publicacoes/saude_brasil_2018_analise_situacao_saude_doencas_agravos_cronicos_desafios_perspectivas.pdf.

[2] Fisher M, Garcia JH. Evolving stroke and the ischemic penumbra. Neurology. 1996;47:884–888.

[3] Nguyen TTM, Kruyt ND, Pierik JGJ, et al. Stroke patient's alarm choice: General practitioner or emergency medical services. Acta Neurologica Scandinavica. 2021;143:164–170.

[4] Markus HS, Brainin M. COVID-19 and stroke—A global World Stroke Organization perspective. International Journal of Stroke. 2020;15:361–364.

[5] Rojas E, Munoz-Gama J, Sepúlveda M, et al. Process mining in healthcare: A literature review. Journal of Biomedical Informatics. 2016;61:224–236.

[6] Leandro G S, Moro C, Miura DY, et al. Process mining leveraging the analysis of patient journey and outcomes: Stroke assistance during the Covid-19 pandemic. Knowledge Management & E-Learning: An International Journal. 2021;421–437.

[7] van der Aalst W, Adriansyah A, de Medeiros AKA, et al. Process Mining Manifesto. 2012. p. 169–194.

[8] Baggio JAO, Santos-Pontelli TEG, Cougo-Pinto PT, et al. Validation of a Structured Interview for Telephone Assessment of the Modified Rankin Scale in Brazilian Stroke Patients. Cerebrovascular Diseases. 2014;38:297–301.

[9] Günther CW, Rozinat A. Disco: discover your processes. In: Lohmann N, Moser S, editors. Proceedings of the Demonstration Track of the 10th International Conference on Business Process Management (BPM 2012). CEUR-WS.org; 2012. p. 40–44.

[10] Jagadeesh Chandra Bose RP, Aalst van der WMP. Context aware trace clustering : towards improving process mining results. Proceedings of the Ninth SIAM International Conference on Data Mining. United States: Society for Industrial and Applied Mathematics (SIAM); 2009. p. 401–412.

[11] Teo K-C, Leung WCY, Wong Y-K, et al. Delays in Stroke Onset to Hospital Arrival Time During COVID-19. Stroke. 2020;51:2228–2231.

[12] TeuschI Y, Brainin M. Stroke Education: Discrepancies among Factors Influencing Prehospital Delay and Stroke Knowledge. International Journal of Stroke. 2010;5:187–208.

Challenges of Trustable AI and Added-Value on Health
B. Séroussi et al. (Eds.)

doi:10.3233/SHTI220395

Automated Diagnosis of Autism Spectrum Disorder Condition Using Shape Based Features Extracted from Brainstem

Vaibhav JAIN[a], Abirami SELVARAJ[a], Rakshit MITTAL[b], Priya RANI[c], Anandh KILPATTU RAMANIHARAN[d] and Jac Fredo AGASTINOSE RONICKOM[a]
[a] *School of Biomedical Engineering, Indian Institute of Technology (BHU), India*
[b] *Telecom Paris, Institut Polytechnique de Paris, Palaiseau 91120, France*
[c] *Applied Artificial Intelligence Institute, Deakin University, Victoria, Australia*
[d] *The University of Alabama at Birmingham, United States of America*

Abstract Alterations to the brainstem can hamper cognitive functioning, including audiovisual and behavioral disintegration, leading to individuals with Autism Spectrum Disorder (ASD) face challenges in social interaction. In this study, a process pipeline for the diagnosis of ASD has been proposed, based on geometrical and Zernike moments features, extracted from the brainstem of ASD subjects. The subjects considered for this study are obtained from publicly available data base ABIDE (300 ASD and 300 typically developing (TD)). Distance regularized level set (DRLSE) method has been used to segment the brainstem region from the midsagittal view of MRI data. Similarity measures were used to validate the segmented images against the ground truth images. Geometrical and Zernike moments features were extracted from the segmented images. The significant features were used to train Support vector machine (SVM) classifier to perform classification between ASD and TD subjects. The similarity results show high matching between DRLSE segmented brainstem and ground truth with high similarity index scores of Pearson Heron-II (PH II) = 0.9740 and Sokal and Sneath-II (SS II) = 0.9727. The SVM classifier achieved 70.53% accuracy to classify ASD and TD subjects. Thus, the process pipeline proposed in this study is able to achieve good accuracy in the classification of ASD subjects.

Keywords Autism spectrum disorder, Brainstem, Level set method, Geometrical features, Zernike moment, Support Vector Machine

1. Introduction

Autism Spectrum Disorder (ASD) is a neurodevelopmental condition characterized by significant challenges in social interaction, communication, and repetitive patterns of behavior which are mainly controlled by brainstem [1]. Anatomical differences in the brainstem region have been identified as an essential biomarker for ASD [2]. Due to its time consuming nature, along with absence of any notable discriminator, the diagnosis of ASD is often delayed [3]. This necessitates the need for a reliable diagnostic technique for ASD. Structural Magnetic Resonance Imaging (sMRI) is one of the most commonly used imaging modality that offers non-invasive methods to screen anatomical anomalies related to neuronal activation. Segmentation of the region of

interest is a daunting task and due to the complexity and resource-intensive nature of manual segmentation, the brainstem is segmented using automated techniques, such as level set methods. Level set algorithms are known to develop irregularities which are generally treated with re-initialization, that stabilizes the level set evolution, however leads to numerical inaccuracies [4]. To avoid the re-initialization problem, distance regularized level set methods (DRLSE) are deployed which use an external energy component and distance regularization term, to force the zero level contours to the prescribed location and maintain its shape [5]. Geometrical [6] and tensor based [7] features are used to infer the shape, size, orientation of brainstem in ASD patients, which indicate the structural variation due to the disease condition. Zernike moments are shape descriptors which can be used as prominent tools as they provide the rotation invariant features [8]. SVM classifier uses various types of kernels to create a decision boundary in the form of a hyperplane and is used to classify ASD and TD using the extracted features.

In this study, the DRLSE method has been used to segment brainstem region from sMRI. The best features from the feature space of segmented regions have been selected and a classification model is trained with optimizations, to aid in the diagnosis of ASD. The rest of the paper includes 3 sections- Section 2 discusses the method used in detail, section 3 describes the results with some discussion about all aspects of the methodology. Finally, section 4 outlines the results with some concluding remarks.

2. Method

2.1. Dataset

The public database Autism Brain Image Data Exchange (ABIDE-I and ABIDE-II) [9] was accessed to obtain sMRI data. The database is classified on the basis of Diagnostic and Statistical Manual of Mental Disorders–Fourth Edition (DSM IV) standards. TD and ASD subjects, 300 each, were carefully chosen to minimize the age difference. The demographics have been shown in Table 1.

Table 1. Participation information of sample set

	Mean ± SD (range)	
	ASD (N=300)	**TD (N=300)**
Males (Females)	260 (40)	223 (77)
Age in years (range)	11.87 ± 2.75 (6.41 - 18)	11.85 ± 2.74 (6.36 - 18.8)
PIQ/FIQ (range)	105.48 ± 17.1 (53 - 149)	111.15 ± 13.75 (62 - 147)

*PIQ/FIQ - Full-scale Intelligence quotient / Performance Intelligence quotient, *SD - Standard Deviation

The mid-sagittal view slices were considered for further processing. A few images with insufficient brainstem regions in the slices were discarded from our analysis.

2.2. Segmentation

The brain stem region is segmented from the sMRI using the DRLSE method [5]. By solving the gradient flow equation (equation (1)), the solution of the energy function is obtained.

$$\frac{\delta f}{\delta t} = \mu\left(d_p|\nabla f|\right) + \lambda\delta(f)\left(g\frac{\nabla f}{|\nabla f|}\right) + \alpha g\delta(f) \quad (1)$$

Here, g represents the gaussian gradient used for edge detection and to avoid the leakage of contours. The parameter values for the level set were set as α=1; μ=0.2; λ=0.1; number of iterations = 15. Trial and error were used to find these sets of parameters. Segmented images were validated against ground truth ones using five similarity measures - Simple Matching, Sokal and Sneath II, Hamann, Rogers and Tanimoto, and Pearson and Heron II, each based on varying criteria.

2.3. Feature Extraction and Classification

The segmented brainstem images were used to extract **Geometrical features** (18), like area, perimeter, eccentricity, orientation, bounding boxes, etc., reported in pixels, coordinates, as well as coefficients. Ratio-metric features (91) [10] were derived from the geometrical ones leading to a total of 109 geometry-based features. **Zernike moments** are mappings of an image onto a set of complex zernike polynomials [8]. Multiple combinations of the order of polynomial (n) and the degree of repetition (m) gave a total of 122 Zernike moments features: 61 Amplitude based (A_{OH}), and 61 recording the Phase angle (Pi_{OH}). The statistical significance of each feature was tested before feeding these into the classifier, by performing normality test to test the nature of distribution, and then t-test and Kruskal Wallis tests for normal and non-normal distribution respectively. SelectKBest method from the scikit-learn library helps choose the optimal feature subset, by ranking features based on F-test, which are then fed as input to the SVM classifier. Scikit-Learn's GridSearchCV returns the best combination of parameters from the input parameter grid, based on the described metric scores. A cross-validation set was used for testing the various parameter combinations.

3. Results

Figure 1(a) and 1(e) represent the mid-sagittal MR brain view of ASD and TD participants respectively. DRLSE methods helped in the evolution of the final contour to properly highlight the brainstem, as shown in figure 1(b) and 1(f). Generalized DRLSE methods detected brain stem boundaries with considerable efficiency. The masked copy of brainstem from the MR images can be observed in figure 1(c) and 1(g) for ASD and TD subjects respectively.

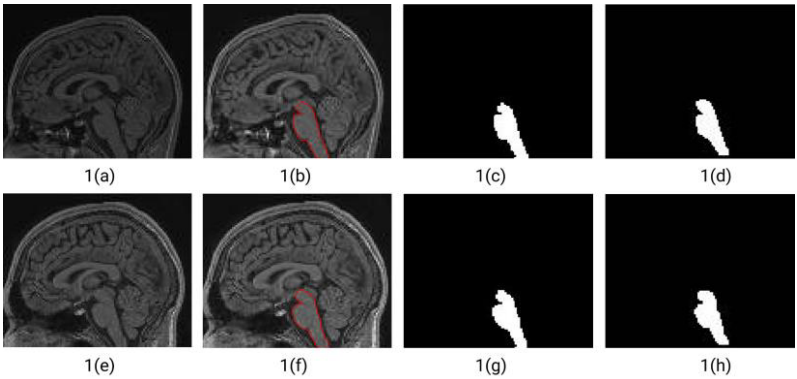


Figure 1. (a, e) Brain Image ASD, TD; corresponding (b, f) Final GDRLSE evolved contour; (c, g) Masked brainstem; (d, h) Ground truth binary image of brainstem

Medical Image Processing, Analysis, and Visualization software (MIPAV) was used to extract the brainstem region from the same dataset of sMRI. These ground truth images obtained can be visualized in figure 1(d) and 1(h) for ASD and TD subjects respectively. The ground truth images served as the validation set for brainstem extracted using the level set method. Validation of segmented brain stem image with the ground truth image using different similarity measures resulted in high correlation (Similarity Index nearer to 1).

The two measures, Pearson and Heron II (PH II), and Sokal and Sneath II (SS II) presented the highest correlation with a mean of 0.9740 and 0.9727 respectively. We calculated the geometrical and Zernike moments features from the DRLSE segmented brain stem images. 59 geometric and 32 Zernike moments features were found to be statistically significant, i.e. had a p-value < 0.05. The top 5 geometrical (including ratio-metric features) and Zernike moment features are presented in table 2.

Table 2.Top five significant geometrical and Zernike moments features

Geometric features	Zernike moments features
Eccentricity/Perimeter	n 7 m 3 A_{OH}
Area/Perimeter	n 12 m 8 A_{OH}
Minor Axis Length	n 17 m 17 A_{OH}
Centroid 2/Convex Area	n 8 m 2 A_{OH}
Convex Area	n 11 m 7 A_{OH}

The Ratio Eccentricity/Perimeter along with Zernike polynomial with the order of moments, n = 7 and the degree of repetition, m = 3 offered highest significance. The order is related to the number of concentric circular divisions, while the degree of repetition shows association with the number of circular sectors. SVM classifier was trained using different types of kernels – Linear, Polynomial, Sigmoidal and RBF. Out of these, the sigmoidal kernel performed best. With one of the five folds being used for testing and 4 being used for training, an accuracy of 70.53 % was reported. The Linear, Polynomial and RBF kernel elicited 58.92%, 57.39%, and 60% accuracy respectively.

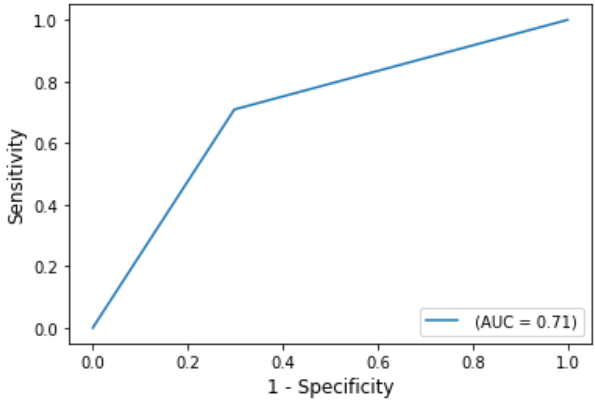


Figure 2. ROC-AUC Curve

The true positive rate (Sensitivity), is the ratio of true positives to the total positive ones, true negative rate (Specificity) is the ratio of true negatives to the total negatives and the F1 Score is the harmonic mean of precision (Positive predictive value) and sensitivity. The sensitivity, specificity and F1 score were reported as 0.6964, 0.7142

and 0.7027 respectively. The Receiver operator characteristics (ROC) curve demonstrating the Area under curve (AUC) can be visualized in figure 2.

4. Conclusions

This study describes the utility of geometrical and Zernike moments features extracted from the brainstem region for diagnosis of ASD. DRLSE based methods efficiently segmented the brainstem region, which was validated by the high similarity index obtained against the ground truth images. The features with the highest significance were obtained and used for classification of ASD from TD providing high accuracy, sensitivity and specificity. The study showed that shape-based features obtained with minimal manual intervention can be used to automate the diagnosis of ASD with relatively higher precision and much lower computational burden. Further work can be done by applying advanced deep learning methods for segmentation and classification of brainstem regions for ASD diagnosis.

5. Acknowledgement

This work was supported by the Science and Engineering Research Board through the Start-up Research Grant (SRG) scheme (SRG/2021/002289).

6. References

[1] Sinclair D, Oranje B, Razak K, Siegel S, Schmid S. Sensory processing in autism spectrum disorders and Fragile X syndrome—From the clinic to animal models. Neuroscience & Biobehavioral Reviews. 2017;76:235-253.
[2] Seif A, Shea C, Schmid S, Stevenson R. A Systematic Review of Brainstem Contributions to Autism Spectrum Disorder. Frontiers in Integrative Neuroscience. 2021;15.
[3] American Psychiatric Association. Diagnostic and Statistical Manual of Mental Disorders. 2013;
[4] Nguyen Duy T, Hino T. An improvement of interface computation of incompressible two-phase flows based on coupling volume of fluid with level-set methods. International Journal of Computational Fluid Dynamics. 2020;34(1):75-89.
[5] Li C, Xu C, Gui C, Fox M. Distance Regularized Level Set Evolution and Its Application to Image Segmentation. IEEE Transactions on Image Processing. 2010;19(12):3243-3254.
[6] Jac Fredo A, Kavitha G, Ramakrishnan S. Segmentation and analysis of brain subcortical regions using regularized multiphase level set in autistic MR images. International Journal of Imaging Systems and Technology. 2014;24(3):256-262.
[7] AR Jac Fredo, G Kavitha, & S Ramakrishnan, Analysis of corpus callosum and its sub- anatomical regions in autistic MR brain images using structure tensors,Second International Conference on Biomedical Systems, Signals and Images, IIT Madras, Chennai, India, 2016
[8] Hwang S, Kim W. A novel approach to the fast computation of Zernike moments. Pattern Recognition. 2006;39(11):2065-2076.
[9] Di Martino A, O'Connor D, Chen B, Alaerts K, Anderson J, Assaf M et al. Enhancing studies of the connectome in autism using the autism brain imaging data exchange II. Scientific Data. 2017;4.
[10] Suganthi M, Madheswaran M. An Improved Medical Decision Support System to Identify the Breast Cancer Using Mammogram. Journal of Medical Systems. 2010;36(1):79-91.

Challenges of Trustable AI and Added-Value on Health
B. Séroussi et al. (Eds.)

doi:10.3233/SHTI220396

Using Explainable Supervised Machine Learning to Predict Burnout in Healthcare Professionals

Karthik ADAPA[a,b,1], Malvika PILLAI[a], Meagan FOSTER[a,b], Nadia CHARGUIA[c] and Lukasz MAZUR[a,b]
[a] *Carolina Health Informatics Program, University of North Carolina (UNC), Chapel Hill, USA*
[b] *Division of Healthcare Engineering, Department of Radiation Oncology, School of Medicine, UNC, Chapel Hill, USA*
[c] *Department of Psychiatry, School of Medicine, UNC, Chapel Hill, USA*

Abstract. Burnout in healthcare professionals (HCPs) is a multi-factorial problem. There are limited studies utilizing machine learning approaches to predict HCPs' burnout during the COVID-19 pandemic. A survey consisting of demographic characteristics and work system factors was administered to 450 HCPs during the pandemic (participation rate: 59.3%). The highest performing machine learning model had an area under the receiver operating curve of 0.81. The eight key features that best predicted burnout are excessive workload, inadequate staffing, administrative burden, professional relationships, organizational culture, values and expectations, intrinsic motivation, and work-life integration. These findings provide evidence for resource allocation and implementation of interventions to reduce HCPs' burnout and improve the quality of care.

Keywords. burnout, healthcare professionals, supervised machine learning

1. Introduction

Burnout is an occupational hazard characterized by emotional exhaustion, depersonalization, and diminished personal achievement. Before the COVID-19 pandemic, 20-40% of healthcare professionals (HCPs) reported severe burnout [1]. The COVID-19 pandemic has further increased HCPs' burnout to levels that pose a threat to maintaining a functioning healthcare workforce [2]. Burnout in HCPs can contribute to low quality of care, impair cognitive processes and lead to patient safety issues including patient harm [3]. Thus, there is an urgent need to examine the key factors contributing to HCPs' burnout during the COVID-19 pandemic.

HCPs' burnout is a complex multi-factorial problem that is often affected by several non-linear factors. The US National Academy of Medicine (NAM) proposed a systems-based framework and identified evidence-based work system factors that contribute to HCPs' burnout [4]. These factors are also further mediated by individual characteristics such as gender, age, and race. However, limited studies have utilized this theoretical model in

[1] Corresponding Author: Karthik ADAPA; E-mail: karthikk@live.unc.edu.

examining the key factors contributing to HCPs' burnout during the COVID-19 pandemic.

In recent times, few studies have applied inductive data-driven methodologies such as supervised machine learning classifiers to predict HCPs' burnout [5]. However, to the best of our knowledge, no previous study has utilized this methodology to examine the role of work system factors and demographic factors in predicting HCPs' burnout during the COVID-19 pandemic. Further, we use feature selection methods to identify key factors that best predict HCPs' burnout to provide evidence for targeting interventions to reduce HCPs' burnout and improve quality of care.

2. Methods

2.1 Data collection and study measures

A composite survey was created to assess the following: demographic factors (clinical position, gender, race, and marital status), burnout using the 2-item Maslach Burnout Inventory (MBI) [6], and severity ratings of 21 evidence-based work system factors based on the NAM's system-based framework [4]. The survey was designed using Qualtrics Online Survey Software and administered to 450 HCPs in oncology, primary care, and surgery departments of a large academic medical center. The survey was administered between November 2020 and May 2021 with a participation rate of 59.3% (Table 1). The study was approved by the UNC-Chapel Hill Institutional Review Board.

The outcome variable is burnout, with emotional exhaustion (EE) (1 to 6) and depersonalization (DP) (1 to 6). An EE and DP summative score >3 correlates best with a more inclusive definition of burnout [6]. However, for this analysis, we considered a score >3 on EE and DP individually as a more restrictive definition of burnout to categorize the HCPs into two classes: with ($\geq$ 3 EE & $\geq$ 3 DP) and without burnout (<3 EE & < 3 DP). The input variables were the 21 work system factors and 4 demographic characteristics.

Table 1. Number and type of survey responses

Feature		Data type	No. (%)
Burnout		Categorical	
	With burnout		105 (30.33%)
	Without burnout		162 (60.67%)
Clinical position		Categorical	
	Physician		70 (26.22%)
	Nurses		89 (33.33%)
	Residents		17 (6.37%)
	Pharmacists		3 (1.12%)
	Non-clinicians		88 (32.96%)
Gender		Categorical	
	Male		42 (15.73%)
	Female		196 (73.41%)
	Non-binary		4 (1.50%)
	Transgender male		4 (1.50%)
	Transgender female		3 (1.12%)
	Prefer to self-describe		3 (1.12%)
	Prefer not to disclose		15 (5.62%)

Race		Categorical	
	Caucasian		180 (67.42%)
	African American		29 (10.86%)
	Latino or Hispanic		5 (1.87%)
	Asian		8 (3.00%)
	Native American		3 (1.12%)
	Native Hawaiian or Pacific Islander		2 (0.75%)
	Other		14 (5.24%)
	Prefer not to disclose		25 (9.36%)
Marital status			
	Single		49 (18.35%)
	Married		161 (60.30%)
	Divorced		14 (5.24%)
	Separated		7 (2.62%)
	Widowed		6 (2.25%)
	Other		6 (2.25%)
	Prefer not to disclose		24 (8.99%)
Work system factors		Ordinal	
Job demands	Excessive workload		264 (98.88%)
	Unmanageable work schedules		261 (97.75%)
	Inadequate staffing		259 (97.00%)
	Time pressure		260 (97.38%)
	Inefficient workflows		258 (96.63%)
	Interruptions and disruptions		257 (96.25%)
	Inadequate technology		256 (95.88%)
	Moral distress		259 (97.00%)
	Patient factors		256 (95.88%)
	Administrative burden		255 (95.51%)
Job resources	Lack of recognition for QI activities		258 (96.63%)
	Lack of dedicated time		257 (96.25%)
	Lack of support for research		255 (95.51%)
	Professional relationships		254 (95.13%)
	Organizational culture		257 (96.25%)
	Physical work environment		254 (95.13%)
	Values and expectations		256 (95.88%)
	Job control		259 (97.00%)
	Intrinsic motivation		253 (94.76%)
	Extrinsic motivation		260 (97.38%)
	Work-life integration		259 (97.00%)

2.2 Feature selection and classification

We selected random forest to predict burnout after weighing the trade-offs between accuracy and interpretability of an array of machine learning methods. The data was split into training (80%) and testing (20%) sets. To avoid overfitting, 5-fold cross-validation (CV) was performed on the training set, where the performance of the model was iteratively evaluated on 20% of the training set. The area under the receiver operating characteristic curve (AUC) was used to evaluate the model. Initially, the RF classifier was trained with factors based on the NAM framework: work system factors & demographic characteristics, and work system factors only. Subsequently, we used chi-square, mutual information, and recursive feature elimination (RFE) to identify attributes that were most important in predicting HCPs' burnout. Features with a mutual information score >0 and chi-square p-value <0.05 were included in the analysis. RFE was done with cross-validation, where features were selected iteratively while

optimizing for AUC performance. After each iteration, the less relevant features were removed, and the key factors that best predicted HCPs' burnout were identified.

3. Results

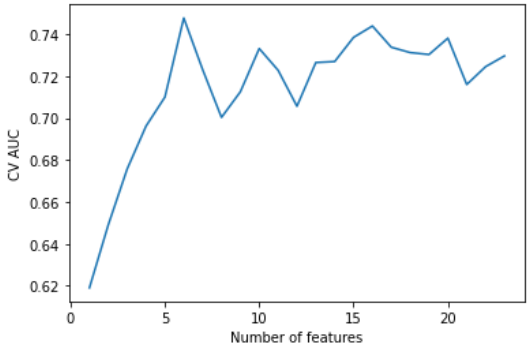


Figure 1. Optimal feature selection using recursive feature elimination

Chi-square showed that only work system factors and race were significantly ($p<0.05$) associated with burnout. RFE with eight features (Figure 1) showed the highest mean CV AUC of 0.755 in comparison to RF models with work system factors & demographic characteristics, work system factors & race, work system factors only and mutual information. In the model testing phase, RFE and work system factors & race showed the highest AUC of 0.811 (Table 1). The eight key features that best predict HCPs' burnout are *inadequate staffing, time pressure, administrative burden, professional relationships, organizational culture, values and expectations, intrinsic motivation, and work-life integration*.

Table 2. Model CV AUC (5-fold and average) and test AUC

Feature selection	**Cross-validation**						**Test**
	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Average	
Work system factors & demographics (all features)	0.642	0.700	0.790	0.714	0.748	0.719	0.798
Work system factors & race	0.632	0.761	0.790	0.717	0.729	0.726	0.811
Work system factors only	0.669	0.787	0.741	0.718	0.746	0.732	0.798
Recursive feature elimination	0.661	0.842	0.805	0.762	0.704	0.755	0.811
Mutual information	0.651	0.822	0.795	0.761	0.714	0.745	0.791

4. Discussion

This study suggests that supervised machine learning methods can be used to examine the role of work system factors and demographic characteristics in predicting HCPs' burnout during the COVID-19 pandemic. Further, this study provides insights into key factors that best predicted HCPs' burnout based on feature relevance and in comparison

to manually curated features from the NAM framework. HCPs are overworked and exhausted after more than a year into the pandemic. Accordingly, job demand factors such as excessive workload, inadequate staffing, and administrative burden seem to better predict HCPs' burnout. Among job resource factors, previous studies [7] have highlighted that deteriorating work-life integration has important consequences on HCPs' well-being. We did not find any demographic characteristics among the eight key predictors of burnout. However, work system factors & race had similar test accuracy as the RFE model and requires further investigation. Thus, preliminary findings from this study could provide evidence to healthcare systems on interventions that can be targeted to reduce HCPs' burnout. In future work, we plan to conduct a sensitivity analysis of the key predictors to strengthen the evidence and improve explainability.

Overall, our study findings are consistent with Nishi et al.'s study [5] that used physician survey data to develop an ensemble of machine learning models with the highest mean AUC of 0.72. Important differences between our study and their study are that they did not use a theoretical model to determine the factors contributing to HCPs burnout, assessed key factors predicting burnout using permutation importance, and did not evaluate performance on a test set. This study has several limitations. First, our results are based on a small sample of HCPs at a single academic medical center. Second, the HCPs who responded to this survey do not represent all medical specialties, groups, and subsets of the healthcare workforce. Thus, our study findings, although promising, cannot be generalized without further investigation.

5. Conclusion

Our study demonstrated that explainable supervised machine learning can be used to predict HCPs' burnout. Among 25 work system and demographic factors, eight factors were identified as the key predictors of HCPs' burnout during the COVID-19 pandemic. Further studies are needed to better understand how machine learning can be used to implement targeted interventions to reduce HCPs' burnout and improve the quality of care.

References

[1] Rotenstein LS, Torre M, Ramos MA, et al. Prevalence of burnout among physicians: A systematic review. *JAMA*. 2018;320(11):1131-1150. https://doi.org/10.1001/jama.2018.12777

[2] Leo, Sabina, Tumolo, Bodini, Ponzini, et al. Burnout among healthcare workers in the COVID 19 era: A review of the existing literature. Front Public Health.2021. https://doi.org/10.3389/fpubh.2021.750529

[3] Mangory, K.Y., Ali, L.Y., Rø, K.I. et al. Effect of burnout among physicians on observed adverse patient outcomes:a literature review. BMC Health Serv Res 21, 369 (2021). https://doi.org/10.1186/s12913-021-06371-x

[4] National Academy of Medicine; Committee on Systems Approaches to Improve Patient Care by Supporting Clinician Well-Being. Taking Action Against Clinician Burnout: A Systems Approach to Professional Well-Being. Washington (DC): National Academies Press (US); 2019 Oct 23. PMID: 31940160.

[5] Nishi M, Yamano M, Matoba S (2021) Prediction of well-being and insight into work-life integration among physician using machine learning approach. https://doi.org/10.1371/journal.pone.0254795

[6] Li-Sauerwine S, Rebillot K, Melamed M, Addo N, Lin M. A 2-Question Summative Score Correlates with the Maslach Burnout Inventory. West J Emerg Med. 2020. 10.5811/westjem.2020.2.45139

[7] Ayar D, Karaman MA, Karaman R. Work-Life Balance and Mental Health Needs of Health Professionals During COVID-19 Pandemic in Turkey. Int J Ment Health Addict. 2021 Nov 24:1-17.

Challenges of Trustable AI and Added-Value on Health
B. Séroussi et al. (Eds.)

doi:10.3233/SHTI220397

An Image Based Object Recognition System for Wound Detection and Classification of Diabetic Foot and Venous Leg Ulcers

Jens HÜSERS[a,1], Maurice MOELLEKEN[b], Mats L. RICHTER[c], Mareike PRZYSUCHA[a], Leila MALIHI[c], Dorothee BUSCH[d], Nina-Alexandra GÖTZ[e], Jan HEGGEMANN[f], Guido HAFER[f], Stefan WIEMEYER[f], Birgit BABITSCH[e], Gunther HEIDEMANN[c], Joachim DISSEMOND[b], Cornelia ERFURT-BERGE[d] and Ursula HÜBNER[a]

[a] *Health Informatics Research Group, Osnabrück University of AS, Germany*
[b] *Department of Dermatology, Venerology and Allergology, University Hospital of Essen, Germany*
[c] *Institute of Cognitive Science, Osnabrück University, Germany*
[d] *Department of Dermatology, University Hospital Erlangen, Friedrich-Alexander University Erlangen-Nürnberg, Germany*
[e] *Department New Public Health, Osnabrück University, Germany*
[f] *Christian Hospital Melle, Niels Stensen Hospitals, Germany*

Abstract. Venous leg ulcers and diabetic foot ulcers are the most common chronic wounds. Their prevalence has been increasing significantly over the last years, consuming scarce care resources. This study aimed to explore the performance of detection and classification algorithms for these types of wounds in images. To this end, algorithms of the YoloV5 family of pre-trained models were applied to 885 images containing at least one of the two wound types. The YoloV5m6 model provided the highest precision (0.942) and a high recall value (0.837). Its mAP_0.5:0.95 was 0.642. While the latter value is comparable to the ones reported in the literature, precision and recall were considerably higher. In conclusion, our results on good wound detection and classification may reveal a path towards (semi-) automated entry of wound information in patient records. To strengthen the trust of clinicians, we are currently incorporating a dashboard where clinicians can check the validity of the predictions against their expertise.

Keywords. Diabetic Foot Ulcer, Venous Leg Ulcer, Wound Care, Artificial Intelligence, Image Classification, Clinical Decision Support System, Health Information Technology

[1] Corresponding Author, Jens Hüsers, Osnabrück University of AS, Health Informatics Research Group, PO Box 1940, 49009 Osnabrück, Germany; E-Mail: j.huesers@hs-osnabrueck.de.

1. Introduction

Chronic wounds are skin and tissue lesions that fail to heal due to different underlying conditions. They have become more frequent in recent years across the world. For example, the British National Health Service witnessed a rise in chronic wound cases from 1.05 million in the year 2014 to 1.58 million in 2018 [1]. In 2018, the most common chronic wound types in the UK were the venous leg ulcer and the diabetic foot ulcer. The increase was over proportionally for both wound types: during the five years, their numbers approximately doubled as the venous leg, and the diabetic foot cases increased by 101% and 93%, respectively. Accordingly, this rise will trigger the future need for wound care and larger consumption of health resources [1].

It is paramount to identify the wound type, i.e., the underlying condition that sustains skin and tissue damage, notably insufficient leg veins for venous leg ulcers [2,3] and diabetic pathophysiology for the diabetic foot ulcer. There are clinical situations where it is the wound image that primarily counts, particularly as more and more images of wounds taken at the point of care become available. Due to the importance of these images, several attempts have been made to support wound assessment and documentation, e.g. [4], and clinical decision making, e.g. [5], through machine learning.

The aim of this study, therefore, is to explore the ability and performance of detection and classification algorithms for diabetic foot ulcers and venous leg ulcers in wound images. This study is embedded in a project on prediction models in wound healing.

2. Methods

We compiled a dataset of wound images from medical records of two specialized wound care centres: the wound care centre of the Christian Hospital Melle and the Department of Dermatology, Venerology and Allergology of the University Hospital Essen, both situated in Germany. All images were part of the wound record, taken during encounters in routine care. The wound type and the images were derived from the medical records, building the ground truth for the classification. The selected images showed at least one diabetic foot ulcer or venous leg ulcer but not both types in one image. A clinician experienced in wound care annotated the type and the corresponding bounding box that located the wound in the image. Afterwards, the annotation was subsequently checked by a second clinician. The final dataset contained 435 diabetic foot ulcers and 450 venous leg ulcers, culminating in 885 images.

The images from both sites were not standardized regarding the angles and distances of the camera. The wounds on the images had different healing stages and provided diverse wound characteristics, among them signs of maceration, infection, or necrosis. All wounds in the images were free from dressing material, and cream or gel remains but may show rulers, information cards, the clinician's hand, or objects in the background. The predominant number of images represent patients with white skin colour.

We used the models from the YoloV5 family for single-shot detection of the two types of chronic wounds. We selected YoloV5 models as they recently showed the best performance in benchmarks for more general tasks [6]. For this task, we explored four YoloV5 models: YoloV5n6, YoloV5s, YoloV5m6, and YoloV5x6. All models used pre-trained weights based on the MS COCO dataset.

The wound images were randomly assigned to a training and test split with a 90% to 10% ratio. The colour channels were scaled from 24-bits to a range of the closed interval

from 0 to 1. A mere 885 images compose a very small training set for modern deep learning architectures. Therefore, the set was artificially enlarged by HSV-noising (hue, saturation value), rotating, translating, scaling, cropping, flipping along the horizontal and vertical axis, which makes the resulting classifier much more robust against factors such as different viewpoints and noise. Additionally, mosaics of different images were randomly compiled during training so that a training image may contain both wound types (Fig. 1 left). The mosaic technique reduced site-specific characteristics in combination with the augmentation mentioned above.

Standard performance metrics for single-shot object detection were calculated for all models. The model showing the highest mean average precision (mAP) with a confidence threshold of 0.5 and an intersection over union (IoU) of 0.5 was selected as the final model.

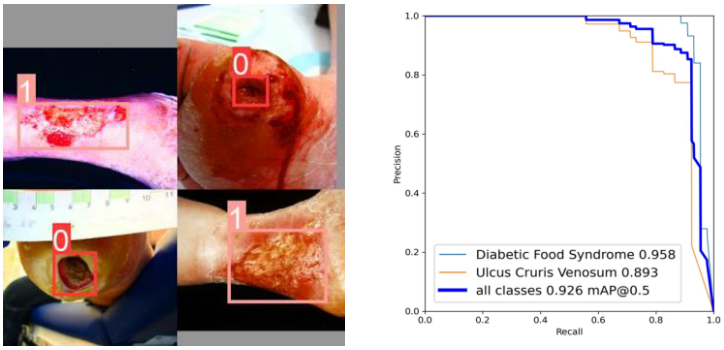


Figure 1 Left: Example of augmented mosaic used for training. In this example, the training image uses four raw images that were randomly augmented and then randomly recompiled into a single training example. The bounding boxes show two diabetic foot ulcers, labelled as 0, and two venous leg ulcers, labelled as 1 **Right**: Recall (x-axis) vs precision (y-axis) curve for the best model (YoloV5m6) for the overall and the class performance with fixed IoU of 0.5 varying confidence levels.

3. Results

Among the 885 images, 789 contained one wound, 85 included two wounds, ten images three wounds, and one image four wounds. The images' average raw width and height were 3374 pixels and 2705 pixels, respectively. All trained models showed convergence and the absence of overfitting. The mean average precision for prediction confidence of 0.5 (mAP_0.5) was at least 0.895 and at most 0.925 when the IoU was set to 0.5. Thus, the models showed low variance and performed similarly for this metric. However, the YoloV5m6, a medium-sized model, had the highest variance for this metric. Compared to the other three models, this model also had the highest precision (0.942) and a high recall value (0.837). Its mAP_0.5:0.95 was 0.642. The model's precision vs. recall curve revealed balanced scores with no sign of overfitting. Furthermore, it shows better predictions for diabetic foot ulcers than venous leg ulcers: the average precision with an IoU of 0.5 (and incrementing confidence levels) was 0.958 and 0.893 (Fig 1 right).

Table 1. Performance metrics per model. The mAP_0.5:0.95 had a fixed confidence at 0.5 and incrementing IoU thresholds for 0.50 to 0.95 in steps of 0.05.

Model	Number of parameters in mio	Recall	Precision	mAP_0.5:0.95	mAP_0.5	F1-score
YoloV5n6	1.9	0.879	0.894	0.638	0.917	0.89
YoloV5s6	7.2	0.908	0.852	0.671	0.920	0.88
YoloV5m6	21.2	0.837	0.942	0.642	0.925	0.89
YoloV5x6	86.7	0.820	0.892	0.652	0.895	0.85

4. Discussion

This study investigated an image-based object recognition system for single-shot wound detection and classification from the YoloV5 family. All models showed satisfying predictive performance, precision, recall, mAP_0.5, and mAP_0.5:0.95. The model with the highest mAP_0.5 metric was the YoloV5m6 model, the medium-sized model with 21.2 million parameters.

This model had a recall (sensitivity) of 0.837, i.e., when a specific wound type is present according to the ground truth, i.e., diabetic foot or venous leg ulcer, the model correctly detects and classifies 83.7% of the present wounds - on average and across both types. Likewise, when a model classifies a detected wound, the fraction of correct detections is 94.2% (precision or positive predictive value). So, when the system predicts and classifies a wound, the probability of a correct prediction is high.

A similar initiative, limited to diabetic foot ulcer detection only [7,8], obtained mAP values in the range of the one in this study. However, they reported lower precision and recall values.

Our results demand cautious interpretation as the images originate from two sites only and, thus, may contain some non-generalizable information. We tried to mitigate this hazard using image augmentation to inflate the dataset and increase its generalizability artificially. Furthermore, our images showed wounds that demand specialized wound care, as both sites focus on the enhanced treatment of complex wound care situations. This may be a limiting factor as less complex wounds and those just on the onset may be underrepresented in the dataset. Therefore, additional images from more sites, including the full spectrum of wound complexity, must be included and tested. Lastly, single-shot detectors may consider background and context information such as anatomical landmarks during detection and classification. For example, when a toe is in the image, odds rise that a diabetic foot ulcer is present rather than a venous leg ulcer. We were able to remedy this effect by compiling mosaic images as training examples to display wounds of both types simultaneously to increase the context independence of the model (Fig 1).

The explainability and transparency of AI algorithms are highly relevant for their safe and reliable application in medicine. In the deep learning architecture used here, it is extremely difficult to extract the features that trigger a specific classification result. However, clinicians can rather easily verify the detection and classification results based on their own medical knowledge, which could help support trust through reproducibility.

It thus could become a potential door opener to accepting clinically more demanding tasks performed automatically. In order to strengthen trust, we are currently enhancing the system with a dashboard where clinicians upload custom images, explore inferences at different confidence thresholds and check the validity of the predictions against their expertise. Furthermore, we are planning to test the findings from the YoloV5 algorithm against other convolutional neural networks for classification and overlay heatmaps that reveal the salient areas.

In conclusion, our results on how to achieve good wound detection and classification may reveal a path towards (semi-) automated entry of wound information in patient records. This functionality may then support clinicians in record keeping and decision support.

Ethical approval

This study was approved by the Ethics Committee of the Osnabrück University of Applied Sciences, Germany (approval no. HSOS/202111/5).

Funding

This study was funded by the German Federal Ministry of Education and Research (BMBF) (grant: 16SV8616).

References

[1] Guest JF, Fuller GW, Vowden P. Cohort study evaluating the burden of wounds to the UK's National Health Service in 2017/2018: update from 2012/2013. BMJ Open. 2020;10(12):e045253.

[2] O'Donnell TF, Passman MA, Marston WA, Ennis WJ, Dalsing M, Kistner RL, et al. Management of venous leg ulcers: clinical practice guidelines of the Society for Vascular Surgery and the American Venous Forum. J Vasc Surg. 2014;60(2 Suppl):3S-59S.

[3] Dissemond J, Kröger K, V ICW e.V. Chronische Wunden: Diagnostik – Therapie – Versorgung. 1st ed. München: Urban & Fischer Verlag/Elsevier GmbH; 2019.

[4] Mukherjee R, Manohar DD, Das DK, Achar A, Mitra A, Chakraborty C. Automated tissue classification framework for reproducible chronic wound assessment. Biomed Res Int. 2014;2014:851582. doi: 10.1155/2014/851582.

[5] Nguyen H, Agu E, Tulu B, Strong D, Mombini H, Pedersen P, et al. Machine learning models for synthesizing actionable care decisions on lower extremity wounds. Smart Health (Amst). 2020;18:100139. doi: 10.1016/j.smhl.2020.100139.

[6] Jocher G, Stoken A, Ayush Chaurasia, Borovec J, NanoCode012, TaoXie, et al. ultralytics/yolov5: v6.0 - YOLOv5n "Nano" models, Roboflow integration, TensorFlow export, OpenCV DNN support [Internet]. Zenodo; 2021 [cited 2022 Jan 22]. Available from: https://zenodo.org/record/5563715

[7] Cassidy B, Reeves ND, Joseph P, Gillespie D, O'Shea C, Rajbhandari S, et al. DFUC2020: Analysis Towards Diabetic Foot Ulcer Detection. Eur Endocrinol. 2021;1(1):5.

[8] Yap MH, Hachiuma R, Alavi A, et al. Deep learning in diabetic foot ulcers detection: A comprehensive evaluation. Comput Biol Med. 2021;135:104596. doi: 10.1016/j.compbiomed.2021.104596.

Challenges of Trustable AI and Added-Value on Health
B. Séroussi et al. (Eds.)

doi:10.3233/SHTI220398

Attitudes and Acceptance Towards Artificial Intelligence in Medical Care

Dana HOLZNER[a,1], Timo APFELBACHER[a,1], Wolfgang RÖDLE[a], Christina SCHÜTTLER[a], Hans-Ulrich PROKOSCH[a], Rafael MIKOLAJCZYK[b], Sarah NEGASH[b], Nadja KARTSCHMIT[b], Iryna MANUILOVA[c], Charlotte BUCH[d], Jana GUNDLACK[e] and Jan CHRISTOPH[a,c,2]

[a] *Department of Medical Informatics, Friedrich-Alexander-Universität Erlangen-Nürnberg, Erlangen, Germany*

[b] *Institute for Medical Epidemiology, Biometrics and Informatics, Interdisciplinary Center for Health Sciences, Faculty of Medicine, Martin-Luther-University Halle-Wittenberg, Halle, Germany*

[c] *Junior Research Group (Bio-)medical Data Science, Faculty of Medicine, Martin-Luther-University Halle-Wittenberg, Halle, Germany*

[d] *Institute for History and Ethics of Medicine, Center for Health Sciences Halle, Medical Faculty, Martin-Luther-University Halle-Wittenberg, Halle, Germany*

[e] *Institute of General Practice and Family Medicine, Center of Health Sciences, Faculty of Medicine, Martin-Luther-University Halle-Wittenberg, Halle, Germany*

Abstract. *Background*: Artificial intelligence (AI) in medicine is a very topical issue. As far as the attitudes and perspectives of the different stakeholders in healthcare are concerned, there is still much to be explored. *Objective*: Our aim was to determine attitudes and aspects towards acceptance of AI applications from the perspective of physicians in university hospitals. *Methods*: We conducted individual exploratory expert interviews. Low fidelity mockups were used to show interviewees potential application areas of AI in clinical care. *Results*: In principle, physicians are open to the use of AI in medical care. However, they are critical of some aspects such as data protection or the lack of explainability of the systems. *Conclusion*: Although some trends in attitudes e.g., on the challenges or benefits of using AI became clear, it is necessary to conduct further research as intended by the subsequent PEAK project.

Keywords. Artificial Intelligence; Physicians; Attitude; Delivery of healthcare; Expert Systems; Machine Learning; Neural Networks, Computer; Computers

1. Introduction

The principle of Artificial Intelligence (AI) goes back to the 1950s when Alan Turing described this type of technology [1]. Although, it is not a new technology, the field of AI seems to be developing very fast nowadays.

[1] These authors contributed equally to this work.

[2] Corresponding Author, Jan Christoph, AG (Bio-)Medical Data Science, Faculty of Medicine, Martin-Luther-University Halle-Wittenberg, Halle, Germany; E-mail: jan.christoph@uk-halle.de.

The PEAK project (prospects for the use and acceptance towards artificial intelligence in healthcare) aims to investigate the attitudes of physicians and patients regarding the use of applications of AI [2]. Current areas of use, possible further areas of use and any gaps in knowledge will be explored. Since there are many different definitions of AI, we settled on one for the PEAK project which encompasses a broad scope: *"the ability to process external data systematically and learn from it to achieve specific goals and tasks. AI involves using machines to simulate human thinking processes and intelligent behaviors, such as thinking, learning, and reasoning, and aims to solve complex problems that can only be solved by experts. As a branch of computer science, the field of AI mainly studies the following contents: machine learning, intelligent robot, natural language understanding, neural network, language recognition, image recognition, and expert system."* [3]. In this sub-study of the PEAK project, physicians were interviewed about their attitudes towards AI.

2. Methods

For the sample description, a total of twelve physicians with experience from different medical disciplines (anesthesiology, cardiovascular surgery, otolaryngology, general medicine, hematology, internal medicine, neurorehabilitation, pediatrics, psychiatry, radiology, radiotherapy, urology, and visceral surgery) aged 28 to 41 years were randomly (regarding to their affinity to AI) chosen and interviewed. There were eight assistant physicians, two specialists and two assistant medical directors with professional experience ranging from 3.5 months to 12 years. The survey covered the entire spectrum from low to high tech affinity.

The interviews were conducted between Nov. 15 and Dec. 14, 2021 by one interviewer, using the "Zoom" web conferencing tool. After oral consent, the conversations were recorded and transcribed by two team members for later analysis. For the evaluation of the expert interviews, the qualitative content analysis according to Kuckartz was applied [4]. For the transcription of the interviews as well as for all evaluation steps and analyses the software MAXQDA was used [5]. In a pretest with four subjects, the interview guide was tested. Since only minor adjustments were made after pretesting and three of the four subjects were physicians, three of the pretest interviews were included in the overall evaluation.

Before conducting the semi structured interviews, a guide was developed. The process of guideline construction was based on the S_2PS_2-method according to Kruse [6]. This consists of the phases "collect, sort, check, delete, subsume". After a brainstorming phase, in which existing literature was incorporated [7-9], unsuitable questions were deleted in discussion with the PEAK team. The remaining questions were subsumed in a structured manner. To arrange the broadest possible focus of potentially addressable topics, the interview guide was filled primarily with qualitative open-ended questions. As an introduction, a query was made about the understanding of the term "Artificial Intelligence" and the current use of AI systems in everyday clinical practice according to perception of the interviewees. In a second step, seven low fidelity mockups i.e., simplified visual representations were used to illustrate possible application scenarios of AI in the clinical context. The mockups served as inspiration for the subsequent general qualitative part, which consisted of acceptance-promoting, acceptance-behind, and topic-specific questions.

For clustering the mockups, four categories were formed, which should cover the widest possible range of usage of AI systems in the medical context: 1. AI in diagnostics, 2. AI in therapy, 3. AI for prognosis/prediction, and 4. AI for process optimization in the hospital. The mockups were illustrated by a PowerPoint presentation that was visible to the participants with screen sharing.

3. Results

3.1. Understanding of the Term Artificial Intelligence and Areas of Application

Regarding the question what the respondents understand of the term Artificial Intelligence the ideas were similar. All twelve interviewees agreed that it is a technical system. Furthermore, about half of the interviewees defined AI as a system that works independently (7/12) and can make decisions on its own (6/12).

Half of the respondents reported not using AI systems in clinical practice to date. However, all but one person had already heard of AI in medical care. After demonstration of the mockups, there were ten people who were using AI applications according to our used definition, mostly in form of medication alerts or voice recognition.

3.2. General Attitudes Towards AI

We asked our twelve interview participants what benefits they believe AI can provide in medical care. For six of the interviewees, the benefit lay in reducing errors in care and thus increasing (patient) safety. Five of the interviewees named the relief of medical work by taking over repetitive or simple tasks, as well as the optimization of processes. One-third of the interviewees saw the benefit of an AI system in improving medical care, structuring data, or saving time, which, according to two interviewees, can in turn be used for patient care.

In contrast to this, five of the interviewees saw seamless integration of systems into existing settings as a major challenge to implementing AI in everyday medical practice, in terms of both program functionality (3/12) and interoperability (2/12). A third of the interviewees saw a risk in relying too much on the systems. Another important point mentioned was the potential for manipulation or misuse of patient data (4/12). In addition, aspects such as the endangerment of the doctor-patient relationship (2/12), the emergence of legal challenges in dealing with AI decisions (2/12), the susceptibility to errors in the development of AI systems recording to the databasis (3/12), and the need for explainability and transparency of the systems, (2/12) were also mentioned.

Further we asked for weaknesses of AI. In this context the interviewees mentioned clinical experience, patient perception and interaction, and human instinct (5/12). In addition, a human is faster than an AI in emergencies. The strengths of AI mentioned included the objectivity of the systems (5/12), matching (1/12) and good presentation (1/12) of data, and that an AI does not miss anything (2/12), has more knowledge (2/12), and does not get tired (2/12).

Regarding the topic of responsibility towards the patient, the physician who uses AI merely as a support for medical decisions still retains responsibility (6/12). The physician, who must question and weigh the AI's decisions, retains the final say (3/12). One subject estimated the responsibility as increased, because a physician has to include an external decision in addition to his own assessment.

Targeting the topic of discrepancy between own medical judgement and AI judgement, most interviewees saw it as an impetus to critically question their own decision and would get to the bottom of the exact cause of the discrepant results (9/12). Half of the respondents would seek the advice of other physician colleagues or supervisors to match opinions with human experience in a discussion.

Moving on to the factors that are influencing trust in AI systems, one of the most frequently mentioned points is the data basis of the systems (7/12). In addition, the transparency of AI systems is of great importance. It was important to six participants that they are informed about the background of the system, such as the nature of the underlying data or knowledge about the development of the programs as well as how the systems work. Other influences on trust included testing of AI systems through clinical trials (3/12).

Regarding the freedom of decision-making for physicians, AI is seen merely as a decision support tool that does not dictate how they should act. More than half (7/12) of the respondents saw no influence on their scope of action. Two feared the subconscious influence of relying more and more on the systems, and three interviewees could imagine AI having a greater influence on decision-making freedom in the future.

4. Discussion

The use of exemplary AI scenarios for the introduction and discussion stimulus of the interviews proved to be useful. Based on lively participation and positive feedback, we could conclude that the subjects were positive about the methodological approach to conducting the interviews. Although the physicians were asked to evaluate AI systems of fields, they were not specialists in, new aspects arose that were thought-provoking and provided valuable input.

The physicians were generally positive about the use of AI in the clinical context. This conclusion was also made in other studies such as Maassen et al. 2021 [9]. An important advantage of the increased use of AI in healthcare is the relief of medical staff and the associated time savings. This saved time would then be available for better patient care, for instance. In the medium to long term, it is entirely conceivable that even this theoretical time savings will fall victim to economic constraints. In contrast, Maassen et al. found that most physicians do not even expect to have more time for patients because of using AI [9].

The interviews revealed that it is important for physicians that AI implementation into everyday life and thus into already existing systems is seamless and without problems. At the same time, care must be taken to ensure that these tools are used wisely in regards of supporting the physicians but not giving them the possibility to rely completely on AI systems´ recommendations. In the end, as Braun et al. stated, it is a human who must make the decision and not AI [10]. Furthermore, according to the physicians, it would be helpful if users of AI could understand why the system makes a decision. We assume that if the attending physician understands why a decision was made, he can also communicate this better to the patient.

A limitation of this study is the small group of twelve experts. Nevertheless, it covers a relatively broad clinical field of eleven medical disciplines. Further research is needed on a much larger group of physicians as well as on patients and the general population. This is planned at a later stage of the study with a total of 800 physicians, 800 patients and 1,000 persons of the general population over the next 2.5 years.

5. Conclusion

Our initial findings show that physicians are generally open to AI in their everyday clinical work. The interviewed physicians see the possibility to get support from AI systems to come to a right decision. Hence, the participants expect that a better decision-making will increase quality in patient care. In addition, according to the physicians´ estimation AI is well suited to assist in some fields of repetitive tasks as it does not get tired and does not need a break. This will be possible especially in areas of factual knowledge matching, such as in the pharmacology interaction check where AI has the advantage of not overseeing something. Furthermore, according to our interviewees AI can also make an important contribution to process and resource optimization which can ultimately lead to time savings. Nevertheless, further studies are needed to cover the current need for more specific information, to clarify specific technical, legal, and ethical issues, and to address the lack of small sample size and small age range in this substudy. This is exactly where the PEAK project will start and therefore, contribute to the research of this topic.

Acknowledgement

We thank all interviewees and all further members of the PEAK team namely Thomas Frese, Jan Schildmann, Daniel Tiller, Susanne Unverzagt and Christoph Weber. PEAK is funded by the Joint Federal Committee (G-BA) under the FKZ 01VSF20017. The present work was performed in fulfillment of the requirements for obtaining the degree "Dr. rer. biol. hum." from the Friedrich-Alexander-Universität Erlangen-Nürnberg.

References

[1] Ramesh AN, Kambhampati C, Monson JRT, Drew PJ. Artificial intelligence in medicine. Ann R Coll Surg Engl. 2004;86(5):334–338.
[2] PEAK – Perspektiven des Einsatzes und Akzeptanz Künstlicher Intelligenz - G-BA Innovationsfonds [Internet]. 2022 [updated 2022 Jan 20; cited 2022 Jan 20]. Available from: https://innovationsfonds.g-ba.de/projekte/versorgungsforschung/peak-perspektiven-deseinsatzes-und-akzeptanz-kuenstlicher-intelligenz.397.
[3] Liu R, Rong Y, Peng Z. A review of medical artificial intelligence. Global Health Journal. 2020; 4 (2): 42–45.
[4] Kuckartz U. Einführung in die computergestützte Analyse qualitativer Daten. 3rd ed. Wiesbaden: VS Verl. für Sozialwiss; 2010. 92 ff. p. (Lehrbuch). ger.
[5] Rädiker S, Kuckartz U. Analyse qualitativer Daten mit MAXQDA. Wiesbaden: Springer Fachmedien Wiesbaden; 2019.
[6] Kruse J. Qualitative Interviewforschung: Ein integrativer Ansatz. Weinheim, Basel: Beltz Juventa; 2014. 227 p. (Grundlagentexte Methoden). ger.
[7] Oh S, Kim JH, Choi S-W, Lee HJ, Hong J, Kwon SH. Physician Confidence in Artificial Intelligence: An Online Mobile Survey. J Med Internet Res. 2019;21(3):e12422.
[8] Laï M-C, Brian M, Mamzer M-F. Perceptions of artificial intelligence in healthcare: findings from a qualitative survey study among actors in France. J Transl Med. 2020;18(1):14.
[9] Maassen O, Fritsch S, Palm J, Deffge S, Kunze J, Marx G, Riedel M, Schuppert A, Bickenbach J. Future Medical Artificial Intelligence Application Requirements and Expectations of Physicians in German University Hospitals: Web-Based Survey. J Med Internet Res. 2021; 23 (3): e26646.
[10] Braun M, Hummel P, Beck S, Dabrock P. Primer on an ethics of AI-based decision support systems in the clinic. J Med Ethics. 2020. doi:10.1136/medethics-2019-105860 Cited in: PubMed; PMID 32245804.

Challenges of Trustable AI and Added-Value on Health
B. Séroussi et al. (Eds.)

doi:10.3233/SHTI220399

Can Artificial Intelligence Enable the Transition to Electric Ambulances?

Emmanouil S. RIGAS[1], Antonis BILLIS and Panagiotis D. BAMIDIS
Medical Physics and Digital Innovation Lab, School of Medicine, Aristotle University of Thessaloniki, Greece

Abstract. The electrification of the transportation sector is seen as a main pathway to reduce CO_2 emissions and mitigate the earth's climate change. Currently, Electric Vehicles (EVs) are entering the market fast. Although EVs have not been used as ambulances yet, the transition to the new type of vehicle is a matter of time. Thus, in this paper we discuss a number of research questions related to the efficient deployment of electric ambulances, focusing on the Artificial Intelligence (AI) point of view and we propose a framework for developing online algorithms that schedule the charging of electric ambulances and their assignment to patients.

Keywords. electric vehicle, electric ambulance, artificial intelligence

1. Introduction

The ever-increasing CO_2 emissions and the consequent greenhouse effect are causing the earth's climate to change. This change creates a major threat for human societies worldwide due to the frequent extreme weather phenomena. Given that the transportation sector is accountable for a substantial portion of greenhouse gas emissions, its electrification, through the use of Electric Vehicles (EVs), is seen as a major pathway to mitigate the climate change. Currently, several EV models have already been introduced to the market and many car manufacturers have set ambitious plans to terminate the production of non-electric vehicles by the end of the current decade.

Despite their many advantages, EVs have certain drawbacks related to their relatively low range and long charging times. To soften these disadvantages, intense research on battery technology is taking place aiming to increase the capacity of the EVs' batteries and, consequently, the autonomy they give to the vehicles. Additionally, charging infrastructure is evolving by the addition of new charging stations, many of them equipped with fast chargers. Now, in order for the EVs to be truly environmentally friendly, they need to charge their batteries using energy produced from renewable energy sources (RES), such as wind or solar [1]. However, these sources are characterized by intermittent production, as they are affected by the weather conditions and time of the day. Thus, the charging of the EVs, as part of the so-called Grid-to-Vehicle (G2V) schemes needs to be scheduled taking into consideration not only the drivers' needs, the available chargers and the network constraints, but it must also

[1] Corresponding Author, Emmanouil Rigas, Medical Physics and Digital Innovation Laboratory, School of Medicine, Faculty of Health Sciences, Aristotle University of Thessaloniki (AUTH), 541 24 Thessaloniki, Greece; E-mail: erigas@auth.gr.

maximize the utilization of the available renewable energy. Otherwise, the need to use EVs at first place would come into question. In this vein, EVs can assist by acting as temporal storage devices when not in use as part of the so-called Vehicle-to-Grid (V2G) schemes [2].

It becomes evident that the management of EVs involves several heterogeneous entities (e.g., EV owners, charging stations), each one having different needs and objectives (e.g., charging the vehicle within a certain deadline, maximizing the profit) which operate in a highly dynamic environment (e.g., charging demand, energy production from RES). Artificial Intelligence algorithms and techniques have already been proven highly efficient in such a domain [3]. For example, Franco et al. [4] propose mathematical programming techniques to schedule EV charging also considering the balance of the electricity network. Additionally, Kadri et al. [5] focus on the placement of the EV charging stations, which is highly important for the efficient charging of the EVs, and they propose both an exact solution using mathematical programming as well as an approximate one using a genetic algorithm. Moreover, Gerding et al. [6] propose mechanism design techniques for the scheduling and pricing of EVs' charging where the vehicles and the stations are considered as intelligent agents. At the same time, Shahriar et al. [7] propose data-driven tools, as well as machine learning (ML) algorithms to learn EVs' charging behavior and improve charging scheduling algorithms. Finally, Quddus et al. [8] propose a solution that is based on Sample Average Approximation with an enhanced Progressive Hedging algorithm in order to expand charging stations by integrating RES and V2G.

Given the progress of the EVs' sector, a crucial question that raises is whether such vehicles can be used as emergency response vehicles, and in particular as ambulances, without compromising their ability to serve people in need. To the best of our knowledge, to date electric ambulances do not operate. However, given the gradual electrification of the transportation sector, eventually ambulances will have to become electric as well. Indeed, UK's NHS is planning to introduce electric ambulances in the near future.[2] Thus, it is important to outline the path that needs to be taken, so as electric ambulances to become a reality and to efficiently replace the conventional ones. In what follows we describe a set of research questions related to the wide deployment of electric ambulances and we explain how AI can assist and, to some extent, enable this deployment.

2. Methods

RQ1: Given the plethora of AI-based algorithms that manage EV activities existing in the literature, which modifications and extensions are needed in order to make these algorithms applicable to electric ambulances? The management of fleets of electric ambulances has two significant differentiations compared to the management of EVs for other domains: 1) The algorithms need to operate in an online fashion and calculate high quality solutions fast as the requests arrive dynamically and 2) the vehicles need to have the highest possible efficiency in terms of the time needed for an ambulance to reach each patient. The existing algorithms schedule EV charging based on the availability of chargers and the spatial and temporal constraints imposed by the users. They operate either offline [9], where the requests are collected in advance and an (optimal) charging schedule is calculated, or online [10] where the requests are collected in a dynamic

[2] https://www.electrive.com/2021/08/09/nhs-procures-fleet-of-electric-ambulances

manner as these are communicated by the drivers. In contrast, when EVs act as ambulances knowing the demand in advance is impossible. Thus, the effort needs to focus in developing online algorithms which will be able to calculate solutions (i.e., assignment of ambulances to each patient request) fast, while utilizing the ambulances' fleet efficiently to minimize delays. In this vein, greedy algorithms with heuristic search can be used, while meta-heuristic approaches can further improve the calculated solutions. Such algorithms can step upon the state of the art algorithms in the EVs' domain. Additionally, previous work on conventional ambulance scheduling such as the work by Erdogan et al. [11] which uses a tabu search algorithm, or the work by Zhen et al. [12] which uses decision rules can be utilized. Moreover, the available charging infrastructure needs to be optimally utilized. In this vein, emphasis should be given on machine learning algorithms and techniques that can predict future demand, in terms of patients' requests, based on historical data, as well as, healthcare service demands as a result of term of the year periodicity or acute events or anomalies (catastrophes and/or pandemics etc).

RQ2: To what extend can electric ambulances utilize renewable sources without sacrificing their ability to transport patients? As discussed before, the utilization of energy from RES is of utmost importance in order the EVs to achieve their purpose and mitigate the greenhouse effect. However, when these vehicles act as ambulances, RES cannot be a priority. It is crucial to have as many ambulances as possible with high battery charge, in order to be able to respond to emergency situations where many ambulances will be needed simultaneously. Thus, the charging scheduling algorithms should consider the use of RES to the extent this does not have a negative effect in the availability of the vehicles. In this context, the use of machine learning techniques that have the ability to predict [13] future production from RES is important.

RQ3: What types and volume of data are available for the algorithms' design and evaluation? Moreover, are these data openly available? In the previous two research questions we outlined the importance of machine learning algorithms and techniques to predict demand for ambulance service and RES production. These algorithms demand data in order to be evaluated and fine-tuned. Thus, the question of the existence and the availability of real-world or realistic synthetic data arises, along with the question of the potential construction of synthetic datasets in order to alleviate potential data scarcity. Last but not least, demands related to open (linked) data availability become apparent.

RQ4: What takes for communities to trust electric ambulances? Public opinion on complex scientific topics can have a big impact on sectors such as the EVs one and to realize the benefits that autonomous systems can provide, they need to be trustworthy by design [14]. Electric ambulances might face a lack of trust from stakeholders and the community in general when it comes to replacing conventional ambulances. This lack of trust might be related both to the technical part of these vehicles, which is out of the scope of this work, and to the AI algorithms that would be used to manage and schedule the electric ambulances. Regarding the latter, the notion of explainable AI [15] comes into place. By using explainable AI techniques, the decisions taken, or the predictions made by the algorithms are explained to the stakeholders in a structured and convincing manner increasing in this way the acceptance of the new technology. To this end, contemporary approaches related to co-creation of solutions with their users and (urban) Living Lab processes [16] involving the whole value of stakeholders may be pivotal.

3. Results

Based on the RQs presented previously, we outline a proposed framework for the scheduling of the charging of electric ambulances and their assignment to patients' requests. This framework is depicted in Figure 1. In detail, initially the problem is mathematically formulated, potentially taking as input experts' opinion collected through participatory design sessions conducted through living labs. Then, the available historical data need to be collected and analyzed using ML algorithms in order to calculate predictions regarding both the future energy production from the available RES and the future patients' demand. Once these predictions are calculated, the focus needs to move to the design of the scheduling algorithm. In this vein and as has been discussed earlier, the developed algorithm needs to have very low execution time in order to calculate both the charging schedule and the assignment of the electric ambulances to the patients fast. The predictions calculated by the ML algorithms should be utilized in order to enhance the performance of the online scheduling algorithm. Regarding the ML and scheduling algorithms, explainability needs to be incorporated in order to assist their acceptance by the stakeholders. Finally, the algorithms need to be evaluated and their performance verified. In this case, an equivalent offline exact solution can be developed, to operate as a theoretical benchmark for the online algorithm. The offline algorithm will collect all the data for a time period (e.g., a day) and calculate the optimal schedule (i.e., the optimal schedule in case we could know all the demand and supply in advance). If the results of the evaluation are considered satisfactory, the procedure terminates, otherwise the algorithms need to be reworked and re-evaluated.

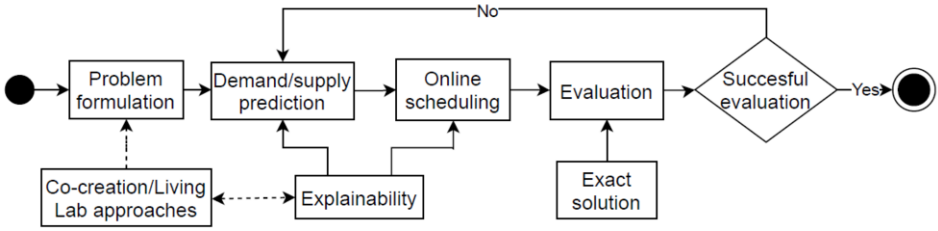


Figure 1. Overview of the proposed framework

4. Discussion and Conclusions

Electric vehicles are entering our lives fast and despite some disadvantages, they will probably dominate the personal transportation market in the next few years. Although electric vehicles have not been widely used as ambulances yet, it is crucial to investigate what are the necessary steps towards making electric ambulances a reality in the years to come. In this paper, once we introduced the concept of EVs and we described the state of the art, we outlined a set of research questions related to the deployment of electric ambulances, from the perspective of AI, and we proposed a general framework for the development of algorithms that can calculate the charging schedule and the assignment of electric ambulances to patients. Although we have not included an exhaustive list of requirements and solutions, we argue that this framework can act as a pathway for future research in this field.

Concluding, for the electric ambulances to become a reality, apart from the advancements in the engineering domain, the development of fast and efficient online scheduling algorithms accompanied with machine learning algorithms able to accurately predict future demand and supply is a key point to achieve a gradual transition from conventional ambulances to electric ones. In this context, AI can have a central role.

Acknowledgement

This project has been partially supported by the HosmartAI project – "Hospital Smart development based on AI", which has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 101016834

References

[1] Sujitha N, Krithiga S. RES based EV battery charging system: A review. Renewable and Sustainable Energy Reviews. 2017;75:978-88.
[2] Kempton W, Tomic J. Vehicle-to-grid power implementation: From stabilizing the grid to supporting large-scale renewable energy. Journal of Power Sources. 2005;144(1):280 294.
[3] Rigas ES, Ramchurn SD, Bassiliades N. Managing Electric Vehicles in the Smart Grid Using Artificial Intelligence: A Survey. Intelligent Transportation Systems, IEEE Transactions on. 2015 Aug;16(4):1619-35
[4] Franco JF, Rider MJ, Romero R. A Mixed-Integer Linear Programming Model for the Electric Vehicle Charging Coordination Problem in Unbalanced Electrical Distribution Systems. IEEE Transactions on Smart Grid. 2015;6(5):2200-10
[5] Kadri AA, Perrouault R, Boujelben MK, Gicquel C. A multi-stage stochastic integer program- ming approach for locating electric vehicle charging stations. Computers Operations Research. 2020;117:104888.
[6] Gerding EH, Stein S, Robu V, Zhao D, Jennings NR. Two-sided online markets for electric vehicle charging. In: Proceedings of the 2013 international conference on Autonomous agents and multi-agent systems; 2013. p. 989-96.
[7] Shahriar S, Al-Ali AR, Osman AH, Dhou S, Nijim M. Prediction of EV Charging Behavior Using Machine Learning. IEEE Access. 2021;9:111576-86.
[8] Quddus MA, Kabli M, Marufuzzaman M. Modeling electric vehicle charging station expansion with an integration of renewable energy and Vehicle-to-Grid sources. Transportation Research Part E: Logistics and Transportation Review. 2019;128:251-79.
[9] Rigas ES, Gerding E, Stein S, Ramchurn SD, Bassiliades N. Mechanism Design for Efficient Allocation of Electric Vehicles to Charging Stations. In: 11th Hellenic Conference on Artificial Intelligence. SETN 2020. New York, NY, USA: Association for Computing Machinery; 2020. p. 10-5.
[10] Alinia B, Hajiesmaili MH, Crespi N. Online EV Charging Scheduling With On-Arrival Commitment. IEEE Transactions on Intelligent Transportation Systems. 2019;20(12):4524-37.
[11] Erdogan G, Erkut E, Ingolfsson A, Laporte G. Scheduling ambulance crews for maximum coverage. Journal of the Operational Research Society. 2010;61(4):543-50.
[12] Zhen L, Sheng S, Xie Z, Wang K. Decision rules for ambulance scheduling decision support systems. Applied Soft Computing. 2015;26:350-6.
[13] Sweeney C, Bessa RJ, Browell J, Pinson P. The future of forecasting for renewable energy. Wiley Interdisciplinary Reviews: Energy and Environment. 2020;9(2):e365.
[14] Ramchurn SD, Stein S, Jennings NR. Trustworthy human-AI partnerships. Iscience. 2021;24(8):102891.
[15] Lundberg SM, Erion G, Chen H, DeGrave A, Prutkin JM, Nair B, et al. From local explanations to global understanding with explainable AI for trees. Nature machine intelligence. 2020;2(1):56-67.
[16] Santonen T, Petsani D, Julin M, Garschall M, Kropf J, Van der Auwera V, et al. Cocreating a Harmonized Living Lab for Big Data–Driven Hybrid Persona Development: Protocol for Cocreating, Testing, and Seeking Consensus. JMIR Research Protocols. 2022;11(1):e34567

Challenges of Trustable AI and Added-Value on Health
B. Séroussi et al. (Eds.)

doi:10.3233/SHTI220400

Using Machine Learning and Deep Learning Methods to Predict the Complexity of Breast Cancer Cases

Akram REDJDAL[a,1], Jacques BOUAUD[a] , Joseph GLIGOROV[b,c] and Brigitte SEROUSSI[a,c,d]
[a]*Sorbonne Université, Université Sorbonne Paris Nord, INSERM, Laboratoire d'Informatique Médicale et d'Ingénierie des connaissances en e-Santé, LIMICS, F-75006 Paris, France*
[b]*Sorbonne Université, Institut Universitaire de Cancérologie, Paris, France*
[c]*AP-HP, Hôpital Tenon, Paris, France*
[d]*APREC, Paris, France*

Abstract. In many countries, the management of cancer patients must be discussed in multidisciplinary tumor boards (MTBs). These meetings have been introduced to provide a collaborative and multidisciplinary approach to cancer care. However, the benefits of MTBs are now being challenged because there are a lot of cases and not enough time to discuss all the of them. During the evaluation of the guideline-based clinical decision support system (CDSS) of the DESIREE project, we found that for some clinical cases, the system did not produce recommendations. We assumed that these cases were complex clinical cases and needed deeper MTB discussions. In this work, we trained and tested several machine learning and deep learning algorithms on a labelled sample of 298 breast cancer patient summaries, to predict the complexity of a breast cancer clinical case. XGboost and multi-layer perceptron were the models with the best result, with an F1 score of 83%.

Keywords. Supervised Machine Learning, Deep learning, Binary classification, Breast Cancer.

1. Introduction

In most countries, therapeutic decisions for cancer patients have currently to be discussed in multidisciplinary tumor boards (MTBs). These meetings have been introduced to provide a collaborative and multidisciplinary approach to cancer care, bringing together surgery, oncology, radiology, and pathology specialists to optimize the decision-making process, and offer the best management to cancer patients. The benefits of MTBs, which have long been taken for granted, are recently being challenged. Positive outcomes from MTBs depend on the presence of qualified experts,

[1]Corresponding Author, Akram REDJDAL; E-mail: redjdalakram300@gmail.com.
This work has been financed by a doctoral grant for AR from the University Institute of Health Engineering (IUIS, Sorbonne University, Paris, France) and received the support of AP-HP health data warehouse

good preparation of patient cases, efficient leadership, time management, and good interactions among MTB clinicians [1].

DESIREE is a European project which aimed at developing web-based services for the management of primary breast cancer by MTBs. During the evaluation of the guideline-based clinical decision support system (CDSS) of DESIREE, we found that for some patient cases the system did not provide any therapeutic proposals or provided recommendations which were not followed by MTB clinicians [2]. These clinical cases were considered as complex cases, and we assume that such cases are not correctly handled by guidelines and need deeper MTB discussion.

Our aim is to optimize the organization of MTBs by identifying complex cases prior to MTBs so that the way clinical cases are discussed could be adapted to allow for a longer discussion for complex cases. We consider the problem of assessing the complexity of a breast cancer clinical case as a binary classification task. We used machine and deep learning models and assessed the performance of such approaches to learn the complexity of breast cancer cases on the basis of breast cancer patient summaries (BCPSs) which are textual documents summarizing the description of the patient condition used as the basis for MTB discussion.

2. Material and Methods

2.1. Breast cancer patient summaries

Breast cancer patient summaries (BCPS) are free text documents used during MTBs. They provide a portrait of a patient case with all relevant information clinicians need to know to make the best patient-specific therapeutic decision. BCPSs contain different types of information, but usually have a common structure with the reason for presentation, tumor type, biometric data, patient personal history, patient family history, TNM staging, etc. Assistance Publique-Hôpitaux de Paris (AP-HP) is the greatest healthcare public university hospital with 39 sites. AP-HP collects patient data, with their consent, in a data warehouse from which BCPSs used in this study were collected.

2.2. Extraction of structured data from BCPSs

We previously used semantic annotators to extract structured data from textual BCPSs [3]. ECMT and SIFR [4] are two annotators that work for the French language. To take advantage of two annotators widely used in the biomedical field for the English language, MetaMap and cTAKES [5], we automatically translated French BCPSs in English. Since BCPSs are textual documents containing a lot of abbreviations, acronyms, and specialized terms related to the oncological field (e.g., “Her2”, “IRM”, “TEP”), a first step was to disambiguate the texts. To solve this issue, we created a local dictionary with medical acronyms and their expansion based on online available dictionaries. Then, we replaced acronyms in BCPSs by their expansion, and finally used the pre-trained Opus-MT translation model (https://opus.nlpl.eu). As a result, all BCPSs were available in both French and English. As a previous work [3] concluded that the application of all four annotators gave the best set of annotations, including complexity-related concepts, we executed the four annotators, and processed the output of each annotator to generate a semantic representation of a BCPS as two vectors, a vector of UMLS concepts (CUI) extracted by SIFR, cTAKES, and MetaMap. And a

second vector containing the labels of the concepts extracted by ECMT (ECMT does not extract UMLS CUIs). For each concept, we associated information about negation as attached to the concept provided by the annotators (e.g., in "*absence d'adénopathie*", the adenopathy concept was present but identified as negated). Figure 1 depicts the whole sequence implemented for the extraction of structured data from BCPSs.

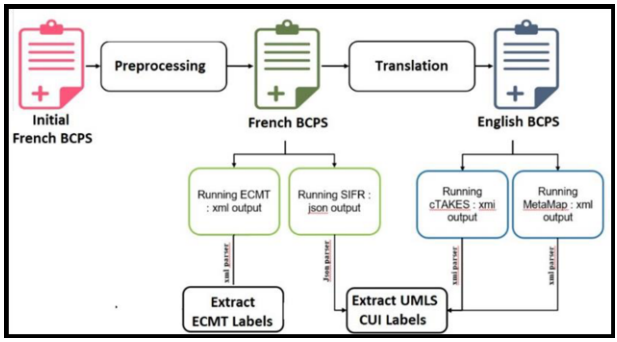


Figure 1. Structured data extraction

2.3. Data preprocessing

In order to get a BCPS representation consumable by all learning models starting from the two vectors obtained from annotators, we converted each BCPS into a row of features that represented the clinical concepts. We included all the labels of the concepts extracted and the value for each feature was 1 if the concept was present, 0 if the concept was not present, and -1 if it was present and negated. We preserved the order of concepts as expressed in a BCPS by using an index column to specify the order in which they appeared in the text.

2.4. Learning algorithms

Three supervised machine learning algorithms were tested (Decision trees, Random forests, XGboost). For deep learning algorithms, we previously tested Multilayer Perceptron (MLP) on a similar task [6], and we wanted to compare MLP performance to models such as Convolutional Neural Networks (CNNs), and Recurrent Neural Networks (RNNs).

The learning dataset was made of a sample of BCPSs which were blindly annotated as "complex" or "non complex" by two expert clinicians. When the experts disagreed, we considered the case as complex.

2.5. Model training pipeline

We trained the six selected models using a k-fold cross-validation strategy, where models' hyperparameters tuning process was executed using Grid Search. Additionally, we selected the most relevant variables for each model using feature selection. Resulting classification models were classically evaluated using precision, recall, and F1-score, both before and after feature selection. Figure 2 displays the whole sequence of treatments for training machine and deep learning model, starting from UMLS CUIs and ECMT labels.

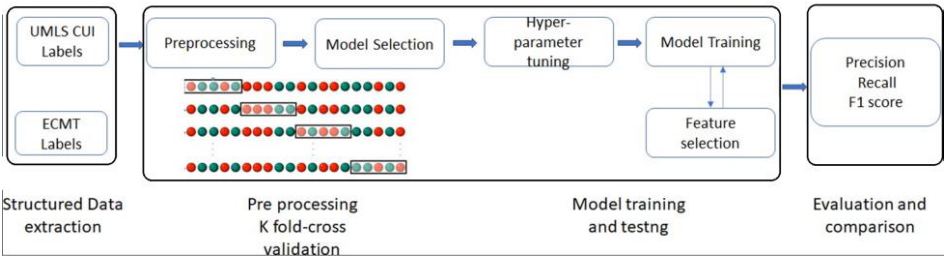


Figure 2. Model training pipeline

3. Results

We worked on a sample of 298 BCPSs corresponding to clinical cases discussed in MTBs between April and July 2021 at the Tenon Hospital in Paris (France). From the 298 BCPSs, we extracted a total number of 10,719 concepts (either negated or not). The presence and negation of these variables in BCPSs were then used as a feature in the inputs. So, for each model, we had a dataset of 10,719 features and 298 patient cases. Following the analytic plan, we generated an initial training dataset and a testing dataset. The training set was divided into 5 splits (stratified-k-fold with k=5) on which we trained each model. All combinations were submitted to Grid Search hyperparamers tunning. The results of each model, with its best parameters are shown in table 1.

Table 1. Evaluation of machine and deep learning algorithms

Models	*Before feature Selection*			*After feature Selection*		
	Precision	Recall	F1	Precision	Recall	F1
XGboost	**0.82**	**0.83**	**0.83**	**0.82**	**0.83**	**0.83**
Random Forest	0.51	0.71	0.59	0.51	0.71	0.59
Decision Trees	0.51	0.71	0.59	0.51	0.71	0.59
MLP	**0.84**	**0.83**	**0.83**	**0.81**	**0.83**	**0.83**
CNN	0.51	0.71	0.59	0.51	0.71	0.59
RNN	0.55	0.65	0.58	0.55	0.65	0.58

4. Discussion and Conclusions

The main goal of this work was to assess the capacity of machine learning and/or deep learning algorithms to predict the complexity of breast cancer clinical cases in order to optimize MTB organization. For instance, a guideline based CDSS recommendation might be followed for non-complex cases, whereas it would be discarded for complex cases.

Starting from a set of textual unstructured BCPSs priorly tagged as complex or not by experts, we were able to extract structured data using semantic annotators and use these structured data as input for learning algorithms. Results showed that XGboost and MLP performed the best with an equal F1 score of 83%. For all models, it is noticeable that feature selection did not improve the results. So, without feature selection, MLP is slightly better than XGboost in terms of precision (0.84 vs 0.82) and comparable in terms of recall (0.83). From this set of models and on this dataset, there was no

evidence of deep learning outperformance over classical machine learning. In our analyses, the MLP model was a multi-layer feed-forward neural network with two layers. More complex network architecture might improve the MLP approach, but this was not tested yet.

Globally, the used algorithms showed good results as compared to previous breast cancer complexity detection research work that we performed in a different context [6]. A similar research work used machine learning from free text to identify breast cancer receptor status with F1 scores reaching 0.89 to 0.92 [7]. Such results are better than ours, but the task looks to be simpler, and the training set used was larger (n=1300). Another work using learning algorithms to predict lung cancer recurrence from EHRs demonstrated a F1 score of 0.72 on a cohort of 2,442 patients [8]. This suggests that our approach is feasible and that our results are acceptable (they could certainly be improved with more annotated BCPSs). This confirms the hypothesis that the use of such algorithms could be useful for this kind of predictions [9]. We wanted to use the same input for both machine learning and deep learning algorithms and the input was a 2-dimension matrix with the presence or absence of concepts retrieved in BCPS texts by semantic annotators. Going further, encoders such as umlsBERT could also be used on BCPs to classify the data in addition to work on a larger dataset, but classical machine learning algorithms could not directly be compared on such a textual input.

References

[1] El Saghir NS, Keating NL, Carlson RW, Khoury KE, Fallowfield L. Tumor boards: optimizing the struc- ture and improving efficiency of multidisci-plinary management of patients with cancer worldwide. Am Soc Clin Oncol Educ Book. 2014:e461-6. doi: 10.14694/EdBookAM.2014.34.e461. PMID:24857140

[2] Redjdal A, Bouaud J, Guezennec G, Gligorov J, Seroussi B. Reusing Decisions Made with One Decision Support System to Assess a Second Decision Support System: Introducing the Notion of Com- plex Cases. Stud Health Technol Inform. 2021 May 27; 281:649-653. doi: 10.3233/SHTI210251. PMID: 34042656.

[3] Redjdal A, Bouaud J, Gligorov J, Seroussi B. Are Semantic Annotators Able to Extract RelevantComplexity-Related Concepts from Clinical Notes? Stud Health Technol Inform. 2021 Nov 18; 287:153- 157. doi: 10.3233/SHTI210836. PMID: 34795101.

[4] Sakji S, Gicquel Q, Pereira S, Kergourlay I, Proux D, Darmoni S, Metzger MH. Evaluation of a French medical multi-terminology indexer for the manual annotation of natural language medical reports of healthcare-associated infections. InMEDINFO 2010 2010 (pp. 252-256). IOS Press.

[5] Reátegui R, Ratté S. Comparison of MetaMap and cTAKES for entity extraction in clinical notes. BMC Med Inform Decis Mak. 2018 Sep 14;18(Suppl 3):74. doi: 10.1186/s12911-018-0654-2. PMID:30255810; PMCID: PMC6157281.

[6] Le Thien MA, Redjdal A, Bouaud J, Seroussi B. Deep Learning, a Not so Magical Problem Solver: A Case Study with Predicting the Complexity of Breast Cancer Cases. Stud Health Technol Inform. 2021 Nov 18; 287:144-148. doi: 10.3233/SHTI210834. PMID: 34795099.

[7] Pironet A, Poirel HA, Tambuyzer T, De Schutter H, van Walle L, Mattheijssens J, et al. Machine Learning-Based Extraction of Breast Cancer Receptor Status from Bilingual Free-Text Pathology Reports. Front Digit Health. 2021 Aug 17 ;3:692077.

[8] Mohamed SK, Walsh B, Timilsina M, Torrente M, Franco F, Provencio M, et al. On Predicting Recurrence in Early Stage Non-Small Cell Lung Cancer. In Proc. AMIA 2021. To appear.

[9] Sekercioglu N, Fu R, Kim SJ, Mitsakakis N. Machine learning for predicting long-term kidney allograft survival: a scoping review. Ir J Med Sci. 2021 May;190(2):807-817. doi: 10.1007/s11845-020-02332-1. Epub 2020 Aug 6. PMID: 32761550.

Challenges of Trustable AI and Added-Value on Health
B. Séroussi et al. (Eds.)

doi:10.3233/SHTI220401

Accelerating High-Dimensional Temporal Modelling Using Graphics Processing Units for Pharmacovigilance Signal Detection on Real-Life Data

Pierre SABATIER[a,b,1] Jean FEYDY[b] and Anne-Sophie JANNOT[a]
[a] *AP-HP.Centre, Université de Paris, Paris, France*
[b] *INSERM INRIA HeKA team, Cordeliers Research Center, Paris, France*

Abstract. Adverse drug reaction is a major public health issue. The increasing availability of medico-administrative databases offers major opportunities to detect real-life pharmacovigilance signals. We have recently adapted a pharmaco-epidemiological method to the large dimension, the WCE (Weigthed Cumulative Exposure) statistical model, which makes it possible to model the temporal relationship between the prescription of a drug and the appearance of a side effect without any a priori hypothesis. Unfortunately, this method faces a computational time problem. The objective of this paper is to describe the implementation of the WCE statistical model using Graphics Processing Unit (GPU) programming as a tool to obtain the spectrum of adverse drug reactions from medico-administrative databases. The process is divided into three steps: pre-processing of care pathways using the Python library Panda, calculation of temporal co-variables using the Python library "KeOps", estimation of the model parameters using the Python library "PyTorch" - standard in deep learning. Programming the WCE method by distributing the heaviest portions (notably spline calculation) on the GPU makes it possible to accelerate the time required for this method by 1000 times using a computer graphics card and up to 10,000 times with a GPU server. This implementation makes it possible to use WCE on all the drugs on the market to study their spectrum of adverse effects, to highlight new vigilance signals and thus to have a global vigilance tool on medico-administrative database. This is a proof of concept for the use of this technology in epidemiology.

Keywords. Graphics Processing Unit, data-driven approach, signal detection, adverse drug reactions.

1. Introduction

Adverse drug reaction is a major public health issue. A recent study by Makary [1], in 2016, places drug-related iatrogeny as the 3rd leading cause of death in the United States. Currently, post-marketing surveillance of drugs is carried out by a pharmacovigilance system, which is based on spontaneous reports from healthcare professionals, pharmaceutical companies and patients. However, this system suffers from under-reporting, with less than 10% of serious adverse reactions being reported [2], and is

[1] Corresponding Author, Pierre Sabatier, Hôpital Européen Georges Pompidou, 20 rue Leblanc, 75015 Paris, France ; E-mail: Pierre.sabatier@aphp.fr

subject to bias due to selective reporting. Most of the reported cases are considered as suspected adverse reactions. The increasing availability of medico-administrative databases offers major opportunities to detect real-life pharmacovigilance signals [3]. However, these databases have mainly been used to confirm or refute potential signals reported [4]. It would be very appropriate to use these databases not to confirm a suspected adverse reaction but to generate adverse reaction signals. The challenge in developing a real-life pharmacovigilance signal detection method presents two difficulties: (a) The choice of a powerful method that can detect a signal among a very large number of drugs and therefore face the problem of multiple tests and therefore the generation of potential false positives, but also face problems of computer resources. (b) Taking into account the temporal heterogeneity of the link between drug intake and adverse effects, which may appear in the short or medium term depending on the pathophysiological mechanism causing the adverse effect.

Over the last decade, several international initiatives have worked on the development of pharmacoepidemiology tools: Mini-Sentinel [5], Observational Medical Outcomes Partnership [6], Pharmaco-epidemiological Research on Outcomes of Therapeutics by a European Consortium [7], Exploring and Understanding Adverse Drug Reactions [8], and the Asian Pharmacoepidemiology Network [9]. The methods compared in these working groups are methods using pharmacovigilance feedback databases (not exhaustive) or so-called "single-candidate" methods, i.e. methods that seek to prove or disprove a hypothesis. These methods do not address our signal detection problem. We have recently adapted a pharmaco-epidemiological method to the large dimension, the WCE (Weigthed Cumulative Exposure) statistical model [10], which makes it possible to model the temporal relationship between the prescription of a drug and the appearance of a side effect without any a priori hypothesis [11], taking into account the temporal heterogeneity described below. Unfortunately, this method faces a computational time problem.

Since 2007, Nvidia has been developing CUDA (originally an acronym for Compute Unified Device Architecture), which is a technology that uses a graphics processing unit (GPU) to perform general purpose computations in place of the processor (CPU), thus speeding up calculations. No methods in epidemiology have been implemented with this technology yet, although it has become widely used in other sectors [12].

The objective of this paper is to describe the implementation of the WCE statistical model using GPU programming as a tool to obtain the spectrum of adverse drug reactions from medico-administrative databases.

2. Methods

2.1. Data-driven WCE

The time-weighted cumulative exposure (WCE) model is used to estimate the cumulative effect of duration, dose and date of prescription of the study drug. The WCE approach estimates the effect of past exposure using a time-weighted sum of all previous exposure events. The weights depend on the time since exposure and the dose. The WCE model is based on a risk function modelled by a spline function and a Cox model with a time-dependent variable to model exposure to the drug of interest. This allows for adverse events that may occur shortly after the start of the study drug or after prolonged exposure. Adjustment variables such as age and gender can be introduced.

The method for estimating the parameters of the WCE model works in two steps. First, the parameters of the risk function are estimated and then the parameters of the Cox model are estimated. The risk function is estimated from the set of patients exposed to the study drug and will be different depending on the drug and/or the event. From this risk function, the WCE model, in a second step, calculates, based on a Cox model, the risk of having the event being exposed to the study drug. The event in our model can be the prescription of another drug, a hospitalization, an illness or a hybrid drug-disease event.

We adapted WCE to an approach without a priori hypothesis by proposing to take into account the multiple tests performed via a bootstrap approach. In the framework of the data-driven WCE extension, the precise consideration of the temporal aspect in the WCE model allows for a good ratio of true positives to false positives in the framework of an approach without a priori hypothesis. The current limitation of this method is its computation time.

In comparison, the most commonly used pharmacovigilance signal detection methods (without a priori hypothesis) are signal disproportionality methods. These methods roughly take into account the temporal aspect of the prescription of the drug (before/after the studied event) and generate many false positives.

2.2. WCE implementation on GPU

The process is divided into three steps:

- Pre-processing of care pathways using the Python library "Panda" - standard for processing tabular data (https://pandas.pydata.org/). Our raw data describes the set of prescriptions associated with a patient population. We extract a list of "times of interest" which correspond to the first intake of selected drugs, used as proxies for adverse side effects.
- Calculation of temporal co-variables using the Python library "KeOps" (https://www.kernel-operations.io). For each time of interest, the prior exposure to the study drug ("suspect") is quantified using correlations with temporal filters - in the example of the original WCE model, these are B-spline functions. The computation of these "temporal co-variables" is efficiently performed using the KeOps library's "block" reduction operations, which distribute the patients over the thousands of cores of our graphics cards.

- Estimation of the model parameters using the Python library "PyTorch" - standard in deep learning (https://pytorch.org/). Given the temporal co-variates computed in the previous step, estimating the maximum likelihood for the Cox/WCE model corresponds to minimizing a convex function with about ten free parameters. We find the solution to this problem using a quasi-Newton method (L-BFGS solver with "strong Wolfe" search), which converges in about ten iterations. We quantify the uncertainty of our estimator using a bootstrap method (N=1,000 to N=10,000 replicates), which is very efficiently implemented with PyTorch using a batch dimension.

We can express the run time of our algorithm as a big O(Its * Drugs * Events * Survivors), where Its is the number of iterations in the convex optimizer for the Cox-PH problem (at most 100), Drugs is the number of drugs of interest, Events is the number of “Death” events - in practice, the prescription of a precise drug that we use as a proxy for

the onset of an adverse effect and Survivors is the typical number of patients that do not develop the adverse effect.

This means that the total run time of our method is at most quadratic in the number of patients (Events and Survivors are both smaller than the total number of patients) and linear in the number of drugs of interest (in practice, 10 to 500). Our GPU implementation does not cut this theoretical time complexity since we still perform the exact same computations, without any approximation. However, they dramatically "improve the constant" by three orders of magnitude. Modern GPUs are made up of thousands of cores.

In practice, the number of patients in our datasets range from 100k to 100M - with an understanding that we always extract subsets that correspond to prescription subgroups. As a consequence, the number of operations required to train our model ranges from about 100 * 100 * 100k * 100k = 10^14 to at most 100 * 100 * 100M * 100M = 10^20 for a hypothetical ubiquitous drug on a very large population. Modern GPUs have a throughput that exceeds 10^12 operations per second: we expect that the run times for our method will range from a few seconds on a single GPU (for small datasets) to a few days on a GPU cluster (for nation-wide studies on very common drugs).

3. Results

In a preliminary work on hydroxychloroquine, the calculation time required for all the prescriptions of 2010 patients over 11 years is 5 days (parallelized calculation on a 2.6 GHz Intel Core i7 6-core processor). To do the same work on 400 ATC classes would take us more than 5 years. This method, as it stands, cannot be used, in real time and on a routine basis, for pharmacovigilance signal detection.

Programming the WCE method by distributing the heaviest portions (notably spline calculation) on the GPU makes it possible to accelerate the time required for this method by 1000 times using a computer graphics card and up to 10,000 times with a GPU server. The calculation to identify the spectrum of side effects associated with a drug on a cohort of about 2000 patients went from 5 days on CPU to 8 minutes (classic PC) or less than 48 seconds (GPU server).

The data driven WCE enabled the identification of eight ATC classes associated with the prescription of HCQ. The most relevant drugs were hydrocortisone, alendronic acid cholecalciferol, valsartan and chlormadinone.

The experimental dataset composed of five tables: patient table with 2010 rows, 3 columns (id_patient, age, sex), size 775KB ; hydroxychloroquine prescription: 50,425 rows, 3 columns (id_patient, date, ATC_classes), size 1.4MB ; all drug prescriptions: 1,175.507 rows, 3 columns (id_patient, date, ATC_classes), size 36.2MB ; all medical procedures: 190,712 rows, 3 columns (id_patient, date, Code_procedures), size 4.8MB ; all medical diagnoses: 48,470 rows, 3 columns (id_patient, date, Code_CIM10), size 1.5MB.

4. Discussion

This paper proposed an implementation of the WCE method using GPU programming allowing to considerably accelerate it. This makes it possible to use this method on all the drugs on the market to study their spectrum of adverse effects, to highlight new

vigilance signals and thus to have a global vigilance tool on medico-administrative database. This is a proof of concept for the use of this technology in epidemiology.

Drug consumption data is remarkably compact and fit without any problem in the main "RAM" memory of our GPUs, which have ~24Gb of usable space. As a consequence, disk latency is much less of a problem than for e.g. medical imaging studies. Going forward, we intend to scale up to full nation-wide studies (~70M patients for the full French population, at most ~500M for the US or the European Union), with full files that may weigh up to 100Gb. These will fit without problem in the RAM of our computers (124Gb) and may be cut into smaller pieces (~1Gb each) to be loaded sequentially on the GPU.

Beyond scaling issues alone, we will seek to extend the WCE model to meet the needs of our pharmaco-vigilance analysis. In particular, we will focus on:

- Patient age management. We will seek to compare standard approaches to Cox/WCE models (which treat age as an additional covariate) with an interval censoring method (already supported by our Python prototype on the GPU).
- Handling more advanced assumptions about the shape of risk functions. Beyond B-Splines (piecewise polynomial), we will evaluate the influence of positivity or convexity assumptions on the relevance of the estimated risk functions.
- Simultaneous management of several drugs. By exploiting the structure of the ATC hierarchical classification, we will seek to go beyond a "parallel" treatment of the different drugs chosen.

References

[1] Makary MA, Daniel M. Medical error—the third leading cause of death in the US. BMJ. 2016 May 3;353:i2139.

[2] Alatawi YM, Hansen RA. Empirical estimation of under-reporting in the U.S. Food and Drug Administration Adverse Event Reporting System (FAERS). Expert Opin Drug Saf. 2017 Jul;16(7):761–7.

[3] Schneeweiss S, Avorn J. A review of uses of health care utilization databases for epidemiologic research on therapeutics. J Clin Epidemiol. 2005 Apr;58(4):323–37.

[4] Weill A, Nguyen P, Labidi M, Cadier B, Passeri T, Duranteau L, et al. Use of high dose cyproterone acetate and risk of intracranial meningioma in women: cohort study. BMJ. 2021 Feb 3;372:n37.

[5] Platt R, Carnahan RM, Brown JS, Chrischilles E, Curtis LH, Hennessy S, et al. The U.S. Food and Drug Administration's Mini-Sentinel program: status and direction. Pharmacoepidemiol Drug Saf. 2012

[6] Stang PE, Ryan PB, Racoosin JA, Overhage JM, Hartzema AG, Reich C, et al. Advancing the science for active surveillance: rationale and design for the Observational Medical Outcomes Partnership. Ann Intern Med. 2010 Nov 2;153(9):600–6.

[7] Trifiro G, Fourrier-Reglat A, Sturkenboom MCJM, Díaz Acedo C, Van Der Lei J, EU-ADR Group. The EU-ADR project: preliminary results and perspective. Stud Health Technol Inform. 2009;148:43–9.

[8] Wisniewski AFZ, Bate A, Bousquet C, Brueckner A, Candore G, Juhlin K, et al. Good Signal Detection Practices: Evidence from IMI PROTECT. Drug Saf. 2016;39(6):469–90.

[9] AsPEN collaborators, Andersen M, Bergman U, Choi N-K, Gerhard T, Huang C, et al. The Asian Pharmacoepidemiology Network (AsPEN): promoting multi-national collaboration for pharmacoepidemiologic research in Asia. Pharmacoepidemiol Drug Saf. 2013 Jul;22(7):700–4.

[10] Sylvestre M-P, Beauchamp M-E, Abrahamowicz M. WCE package: weighted cumulative exposure models. :10.

[11] A data-driven pipeline to extract potential side effects through co-prescription analysis: application to a cohort study of 2,010 patients taking hydroxychloroquine with an 11-year follow-up. Available from: https://www.medrxiv.org/content/10.1101/2021.04.14.21255406v1

[12] Feydy J, Glaunès J, Charlier B, Bronstein M. Fast geometric learning with symbolic matrices, NeurIPS 2020.

Challenges of Trustable AI and Added-Value on Health
B. Séroussi et al. (Eds.)

doi:10.3233/SHTI220402

Analysis of Saturation in the Emergency Department: A Data-Driven Queuing Model Using Machine Learning

Adrien WARTELLE[a,b,1], Farah MOURAD-CHEHADE[b], Farouk YALAOUI[b], David LAPLANCHE[b] and Stéphane SANCHEZ[b,c]
[a] *Computer Science and Digital Society Laboratory (LIST3N), Université de Technologie de Troyes, Troyes, France (e-mail: adrien.wartelle@utt.fr, farah.chehade@utt.fr, farouk.yalaoui@utt.fr)*
[b] *Public Health and Performance Department, Centre Hospitalier de Troyes, Troyes, France (e-mail: david.laplache@hcs-sante.fr, stephane.sanchez@hcs-sante.fr,*
[c] *Research on Health University Department – University of Reims Champagne Ardenne, Reims, France*

Abstract. Emergency department is a key component of the health system where the management of crowding situations is crucial to the well-being of patients. This study proposes a new machine learning methodology and a queuing network model to measure and optimize crowding through a congestion indicator, which indicates a real-time level saturation.

Keywords. Emergency department, Saturation, Crowding indicator, Queuing model, Data-driven-model, Machine learning-based forecasting model, Simulation-Optimization.

1. Introduction

Emergency department (ED) is a key component of the health system that acts as a safety net of healthcare for the population around. With its unscheduled nature and its ease to be accessed, it is subject to real-time crowding phenomena with serious consequences on patients' health and on doctor and nurses working conditions (1). Its causes have been related to the input of ED, that is patient arrivals volumes, the throughput with examination services time performances and output with downstream blocking of patients due to shortage of hospitalization beds.

ED crowding is a complex problem not only regarding its causes but also regarding its definition and the way it is quantified as illustrate the many ED crowding scores existing in the literature such as the EWIN, READI, SONET and NEDOCS (2,3). They each try to quantify the stress that patients' healthcare demand put on the ED system in comparison to the availability of resources. Especially, the National Emergency Department Overcrowding Score (NEDOCS) has been designed to correlate the most with the empirical real-time evaluation of ED crowdedness by field experts using a linear regression on 5 key variables (4). However, all these scores do not necessarily better than

[1] Corresponding Author, Adrien WARTELLE; E-mail: adrien.wartelle@utt.fr.

global mean occupation to predict consequences of crowding and the variables they use are not always available in data records (3,5). As such, crowding can be seen mainly as an occupation level problem which is in turn a direct consequence of saturation periods where the rate of patient arrivals is greater than the rate of departures.

Queuing networks models are well suited to model these crowding events as they can capture all these metrics. Their design differs mainly on four criterions which are their granularity, the queuing mechanisms with the type of servers involved, the type of variables modeled to characterize the service processes and the technique of estimation for arrivals and services time or rate variables. The model granularity is linked to the number of stations and patient pathways and can go from one station model (6) to complex detailed models with the distinction of each action and processes (2,7,8). The servers of the stations can represent resources such as beds, nurses, and doctors (8), or be more abstract with a single server or infinite servers (6,9). The services processes can also be seen either as a succession of services times (2,7,8) or as a point process with a variable intensity of service rate (6,9). Their end goal is either to predict future crowding values using data-driven approach (3,6), or to optimize crowding with theoretical model which takes into account the effect of decision variables such as staff planning (8). However, no model exists yet to both forecast and optimize crowding at the same time which could provide better and more robust planification and decision-making strategies that could in-turn be adapted given the crowding context and the time scale considered.

The aim of this study was to propose a new methodology perspective on crowding measurement, forecasting and optimization using data-driven queuing approach as well as an original congestion indicator to measure saturation.

2. Methods

2.1. Design and Definition of congestion

We conduct a large retrospective observational study based on emergency department record during 2017 to 2019. Emergency department is a service system which can be captured by a mathematical model on a discrete space with continuous time using an arrival process $A(t)$ and a departure process $D(t)$ which describe, respectively, the number of arrivals and departures of the system since an initial period $t = 0$. Instead of focusing only on the occupation level $l(t) = A(t) - D(t)$, the present study focuses on an original congestion measure:

$$\rho(t) = \frac{f(A(t) - A(t-h))}{g(D(t) - D(t-h))} \; \forall t \in \mathbb{R}^+, t > h \quad (1)$$

The congestion $\rho(t)$ uses a ratio comparison of arrivals to departures on a time window $[t-h, t]$ and with regulating functions $f(.)$ and $g(.)$ which are both chosen to ensure a local and stable indicator where $\rho(t) > 1$ indicates local system saturation.

2.2. Queuing methodology

To model this new indicator, the stochastic arrival and departure processes $A(t)$ and $D(t)$ need to be fully modelized. To achieve this goal, this study proposes an ED queuing network using a set of elementary stations $i \in I$, with their arrivals $A_i(t)$ and departures

$D_i(t)$ processes, including an external environment indexed by 0 and set of stations $F \in \mathcal{F} \subset \mathcal{P}(I)$. For practical purpose, a discretization of space is introduced with $A([t, t+k]) = A(t+k) - A(t)$ (same for D) and $k \in \mathbb{R}^+$ the length of time interval unit. The arrival processes of each station are connected to the departure of others with transition processes D_{i_1,i_2}, $i_1, i_2 \in I$ and governed by the transition probabilities $p_{i_1,i_2}([t, t+k])$. As illustrated in Figure 1, the considered queuing network is a 7 station ED patient flow model with 2 tracks, a short circuit for patients with low-acuity health problems and a long circuit for the others. The patients follow a classic ED process of nurse triage, doctor initial examination, supplementary examinations and a holding before leaving the ED for home or hospitalization in most cases.

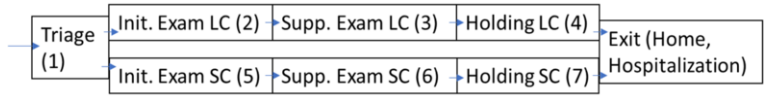


Figure 1. General ED Patient Flow model (LC: Long Circuit, SC: Short Circuit)

The final goal of the queuing approach is to model the distributions $P(A_F([t, t+k] = n), P(D_F([t, t+k] = n), \forall n \in \mathbb{N}, \forall F \in \mathcal{F}, \forall t \in \{t = nk, n \in \mathbb{N}\}$. The proposed modelization considers each of the 7 departure processes of the ED queuing network as Coxian processes described by a local departure rate function whose behavior is inferred using past information on service performances and current information about the number of patients in the ED and the number of triage nurses and doctors in long and short circuits (10). The obtained queuing model can then be simulated evaluate, forecast, and optimize crowding using notably the original congestion measure. This study was performed in compliance with the national legislation regarding epidemiological studies (Declaration N° 2203674v0). Since the study was wholly observational and only used anonymized data (patient names were not recorded), neither ethics approval nor a specific written informed consent from participants were required under French law as a retrospective database study.

3. Results

The data involved for the evaluation of the method come from the ED of Troyes Hospital in Eastern France during the year 2017 to 2019 with a focus on 2019. It is the largest hospital in the Aube Department of France which has a population of 310,000 inhabitants and a medical density of 234.1 physicians per 100,000 inhabitants. In 2018, there were a total of 62,082 ED visits corresponding to an average use rate of 250 to 330 visits per 1,000 inhabitants within the hospital's service area.

3.1. Congestion characteristic of the ED of Troyes

For each 4-hour period of the day of 2019, the Table 1 describes key crowding metrics with the mean 2h-lag number of arrivals (A), the mean 2h-lag number of departures (D), the mean occupation level (L) and the mean 2h-lag congestion ($\bar{\rho}$) (h=2). To keep a stable finite congestion, the regulation functions add one arrival and one departure event.

Table 1. Key ED crowding metrics for Troyes ED in 2019

Hour period	0h-4h	4h-8h	8h-12h	12h-16h	16h-20h	20h-24h	0h-24h
L	40.7	24.4	16.2	26.7	42.7	45.4	32.7
$\bar{\rho}$ (~A/D)	0.9 (~17.2/21.3)	0.6 (~9.0/17.1)	0.9 (~5.4/7.8)	2.4 (~16.0/7.9)	1.6 (~21.8/15.0)	1.0 (~21.0/21.3)	1.2 (~15.1/15.1)

3.2. Queuing model parameters and validity

The first and most important step of the queuing network development which is to model the departure process of each station has been undertaken. The departures rates are currently modeled using logistic regression on the probability of at least one departure on 1 second intervals which is taken as a polynomial function of the variables described in the method section. Figures 2 and 3 show the estimated departures rates per hour of the model for the initial examination stations (EXAMCLI for long circuit and EXAMCCI for short circuit) for short circuit depending on the current number of medical actors (LM) and patients (LP) present.

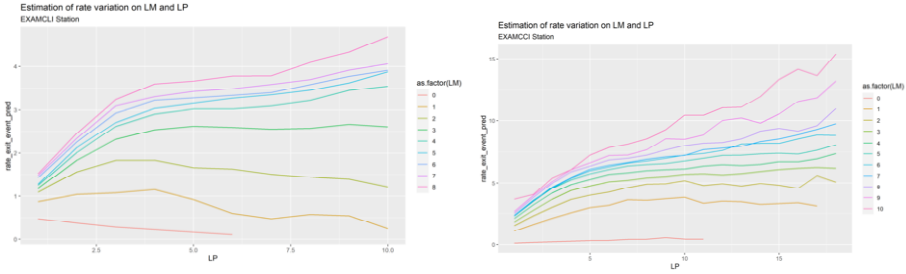


Figure 2 and 3. Estimated rates of departures

Although the forecasting performances of congestion have not yet been evaluated, the models have been checked for the adequation between mean occupation level observed over the year 2019 (L) and the one produced by simulation ($\hat{L}$) of each station independently considering their arrival process as known using the mean of 10 simulation replications.

Table 2. Simulation of ED Troyes 2019 crowding occupation, bias fit

Station	T (1)	I.E LC (2)	S.E LC (3)	H LC (4)	I.E SC (5)	S.E SC (6)	H SC (7)	Sum
Bias: $\hat{L}$ -L	3.3-3 = 0,3 (11.4%)	4.4-4.1=0,3 (8.7%)	11.2-11.2=0 (0.7%)	2.3-2.3=0 (-1.5%)	5.5-5 = 0.5 (9.6%)	5.6-5.7 =0.1 (-1.2%)	1.7-1.5 = 0.2 (14.0%)	34-32.7 = 1.3 (3.7%)

4. Discussion and Conclusion

The queuing model approach for forecasting and optimization of saturation is an original and promising approach. As Table 1 illustrates most of the crowding accumulation ED happens during the beginning of the afternoon (12h-16h) where the mean 2h lag congestion is the highest at 2.4 whereas the peak occupation is attained during the period 20h-24h. The lag of 2 hours has been chosen to give information as local as possible on the saturation as well as to obtain enough events whose mean varies here from 13.2 events (8h-12h) to 42.3 (20h-24h). The results of the queuing models show that they can adequately capture the service performance variations based on the number of patients and of medical actors.

Despite these current promising results, the model still needs to be analyzed further to show its performances to forecast crowding in terms of congestion and occupancy using adequate error metrics and compare them to the literature on forecasting ED occupancy. These forecasting performances are still under encouraging investigation. In the meantime, this study already shows an adequate representation of ED performances and is suitable for simulation-optimization purposes with a 3.7% relative bias error on the total ED occupation, not considering yet the modelization of arrivals in the ED and the transitions models. Furthermore, the queuing design approach can be easily adapted to suit any ED, and even any service system, as long as arrival and departure datetimes are available for forecasting purposes and staff planning or other decision variables can be extracted for optimization. It will form the basis of a management tool to detect future congestion situation in the short and long term and propose solutions toward optimizing it. Staffing planification will form the key strategy of this optimization (11), and will be completed with other complementary strategies. These strategies will consider reorganization of the triage process (12), and the redirection of non-emergent cases (13), toward alternative unscheduled primary care services.

To conclude, our approach explores new methods with current emergency record data to drive healthcare system in real time.

References

[1] Morley C, Unwin M, Peterson GM, Stankovich J, Kinsman L. Emergency department crowding: A systematic review of causes, consequences and solutions. PLoS One [Internet]. 2018;13(8):e0203316.

[2] Ahalt V, Argon NT, Ziya S, Strickler J, Mehrotra A. Comparison of emergency department crowding scores: a discrete-event simulation approach. Health Care Manag Sci. 2018;

[3] Hoot NR, Zhou C, Jones I, Aronsky D. Measuring and Forecasting Emergency Department Crowding in Real Time. Ann Emerg Med [Internet]. 2007 [cited 2021 Jul 2];49(6):747–55.

[4] Weiss SJ, Derlet R, Arndahl J, Ernst AA, Richards J, Fernández-Frankelton M, et al. Estimating the degree of emergency department overcrowding in academic medical centers: results of the National ED Overcrowding Study (NEDOCS). Acad Emerg Med [Internet]. 2004 [cited 2022 Feb 10];11(1):38–50.

[5] Cheng Q, Argon NT, Evans CS, Liu Y, Platts-Mills TF, Ziya S. Forecasting emergency department hourly occupancy using time series analysis. Am J Emerg Med. 2021 Oct 1;48:177–82.

[6] Whitt W, Zhang X. Forecasting arrivals and occupancy levels in an emergency department. Oper Res Heal Care [Internet]. 2019 Jun 1 [cited 2020 Jan 18];21:1–18. Available from: https://www.sciencedirect.com/science/article/abs/pii/S2211692318300407

[7] Bair AE, Song WT, Chen YC, Morris BA. The impact of inpatient boarding on ED efficiency: A discrete-event simulation study. J Med Syst. 2010;

[8] Cochran JK, Roche KT. A multi-class queuing network analysis methodology for improving hospital emergency department performance. Comput Oper Res. 2009;

[9] Sharafat AR, Bayati M. PatientFlowNet: A Deep Learning Approach to Patient Flow Prediction in Emergency Departments. IEEE Access. 2021;9:45552–61.

[10] Vere-Jones, D. J., Daley D. An Introduction to the Theory of Point Processes. An Introduction to the Theory of Point Processes. Springer-Verlag; 2003.

[11] Feng Y-Y, Wu I-C, Chen T-L. Stochastic resource allocation in emergency departments with a multi-objective simulation optimization algorithm. Heal Care Manag Sci 2015 201 [Internet]. 2015 Aug 5 [cited 2021 Oct 18];20(1):55–75. Available from: https://link.springer.com/article/10.1007/s10729-015-9335-1

[12] Chmielewski N, Moretz J. ESI Triage Distribution in U.S. Emergency Departments. Adv Emerg Nurs J [Internet]. 2022 Jan [cited 2022 Jan 31];44(1):46–53. Available from: https://pubmed.ncbi.nlm.nih.gov/35089282/

[13] Tapia AD, Howard JT, Bebo NL, Pfaff JA, Chin EJ, Trueblood WA, et al. A Retrospective Review of Emergency Department Visits That May Be Appropriate for Management in Non-Emergency Settings. Mil Med [Internet]. 2022 Jan 18 [cited 2022 Jan 31]

Challenges of Trustable AI and Added-Value on Health
B. Séroussi et al. (Eds.)

doi:10.3233/SHTI220403

Pretrained Neural Networks Accurately Identify Cancer Recurrence in Medical Record

Hussam Kaka [a], George Michalopoulos [a], Sujan Subendran [a], Kathleen Decker [b], Pascal Lambert [b], Marshall Pitz [b], Harminder Singh [b], Helen Chen [a,1]

[a] *University of Waterloo*
[b] *CancerCare Manitoba*

Abstract. Cancer recurrence is the diagnosis of a second clinical episode of cancer after the first was considered cured. Identifying patients who had experienced cancer recurrence is an important task as it can be used to compare treatment effectiveness, measure recurrence-free survival, and plan and prioritize cancer control resources. We developed BERT-based natural language processing (NLP) contextual models for identifying cancer recurrence incidence and the recurrence time based on the records in progress notes. Using two datasets containing breast and colorectal cancer patients, we demonstrated the advantage of the contextual models over the traditional NLP models by overcoming the laborious and often unscalable tasks of composing keywords in a specific disease domain.

Keywords. Natural language processing, Real-world data, Breast cancer, Colorectal cancer, Cancer recurrence, BERT, ClinicalBioBert architecture

1. Introduction

Breast and colorectal cancer are two of the most common cancer types globally. They are the second and third leading causes of cancer death in North America [4]. Both diseases carry a risk of recurrence after treatment, and recurrent cases need to be flagged expediently for further treatment to increase the chances of survival. Cancer patients tend to have many interactions with clinicians throughout their treatment, resulting in many unstructured medical records that are difficult to search. Consequently, quality control interventions such as identifying patients with recurrence who are not yet on a treatment plan require manual chart abstraction, which is time-consuming and sometimes inaccurate [6]. Moreover, retrieval of information on recurrence becomes difficult during clinical encounters, potentially affecting the patient experience.

Previous approaches for cancer recurrence detection include using treatment codes, such as those for chemotherapy or radiation, detecting breast and lung cancer recurrence [8]. Still, treatment codes can vary by jurisdiction and over time. Other approaches in-

[1]Corresponding Author, Helen Chen, University of Waterloo, School of Public Health Sciences; E-mail: helen.chen@uwaterloo.ca

clude generating a vocabulary of terms specific to the cancer of interest by specialized oncologists, which can then be used for parsing the notes and recurrence detection [5].

In the past decade, the field of natural language processing, which studies information extraction from text, has undergone significant advances, culminating in the development of robust transformer-based models [12]. These models consist of large neural networks that are "pre-trained" on large datasets and then transferred to the task at hand. Bidirectional Encoder Representations from Transformers (BERT)[7] is a transformer-based model which has shown excellent results in many language tasks, including classification and named entity recognition. It is pre-trained on general-purpose English text, and the resulting pre-trained model, BERT-base, can be subsequently used to detect cancer recurrence. A variation of this, ClinicalBioBERT, was trained on medical data to improve its performance on medical tasks [1]. Other published models include BioBERT [10], which is trained on PubMed abstracts and full texts, and UmlsBERT [11], which incorporates information from the Unified Medical Language System (UMLS) [3].

In this study, two models named BERT-base and ClinicalBioBERT, are used to detect cancer recurrence on unstructured medical notes. They are trained and evaluated for the detection of breast cancer recurrence and colorectal cancer recurrence without the need for expert input in the form of specialized vocabulary or decision rules.

2. Material

In this study, we used two cancer datasets (breast and colorectal) which were curated from electronic medical record notes at Cancer Care Manitoba.

Each of these datasets was split to an internal and an external set. The internal dataset was used for hyperparameter tuning and model training, and it consisted of all notes dated between 2004 and 2007. The external dataset was used for final model evaluation, which consisted of notes generated between 2008 and 2012.

Both breast and colorectal cancer datasets demonstrated a skew towards negative instances. Of the 112,285 notes in the breast cancer internal dataset, only 5,082 (4.3%) were positive for cancer recurrence. Of the 116,146 notes in the colorectal cancer internal dataset, only 4,207 (3.6%) were positive. We observed similar trends for the external breast and colorectal cancer datasets. Full dataset characteristics are shown in Table 1.

	Breast Cancer		Colorectal Cancer	
	Internal	External	Internal	External
N (notes)	112,285	78,469	116,146	122,262
Recurrent Cancer	5,082	2,985	4,207	4,245
# patients	897	615	536	589

Table 1. statistics of the datasets

3. Method

We used two pre-training models: BERT-base and ClinicalBioBert. They are identical in network architecture but differ in their pre-training datasets. BERT-base is pre-trained on

general English texts (Books Corpus and English Wikipedia), while ClinicalBioBERT is further trained using biomedical texts (PubMed Abstracts and PubMed Central Papers) and clinical texts (the MIMC-III dataset [9]). Therefore, ClinicalBioBert, with medical context, serves as the target model and BERT-base, with general English pre-training, serves as the control.

BERT models convert input text into tokens before processing. The tokenization process converts input words into a numeric representation based on a provided dictionary. Since the dictionaries are limited in size (approximately 30,000 words in BERT-base), the tokenizer will inevitably encounter dictionary words that are split into subcomponents. Thus, a single word can be represented by one or more tokens. The vocabulary used in this study consisted of 28,996 words and was originally published by Devlin et al. [7] and subsequently used by ClinicalBioBERT. The attention mechanism used by BERT exhibits quadratic runtime complexity with respect to the input length, necessitating a limit to input length. Pre-trained BERT-base and ClinicalBioBERT models used in this study set the input length limit to 512 tokens. Since medical notes can easily exceed this length, a mechanism to shorten notes is required for these models to be usable. We hypothesized that the most important information in medical notes would be included at the beginning or the end of a note. Thus the input document of the model consists of the first 256 and last 256 tokens of each note.

Each breast cancer and the colorectal cancer datasets were divided 90%-10% for training and validation, respectively. A grid search was performed on the learning rate and batch size with 5-fold cross-validation to obtain the optimal hyperparameters for the model. Binary cross-entropy with class weights corresponding to dataset prevalence was used as the loss function. The best-performing model hyperparameters were then selected and used to train a final model for each dataset using 100% of the data. The final models were then evaluated on the external datasets. The selection of the optimal hyperparameters was based on the area under the receiver operating characteristic curve (ROC-AUC).

It should be noted that independent dataset metrics were used for model evaluation. These metrics were: (i) ROC-AUC (ii) sensitivity (iii) specificity (iv) the modified Brier score. Dataset-dependent metrics were used for the assessment of model suitability for clinical use. In this setting, the unbalanced accuracy, positive predictive value (PPV) and the negative predictive value (NPV) were used.

4. Results-Discussion and Conclusion

The results of the 5-fold cross-validation are shown in Table 2. As seen by the ROC-AUC results, ClinicalBioBERT and BERT-base showed very close results, within one standard deviation of each other.

	Breast Cancer Dataset	Colorectal Cancer Datase
ClinicalBioBERT	0.9955 ± 0.0006	0.9921 ± 0.0074
BERT-base	0.9948 ± 0.0014	0.9912 ± 0.0091

Table 2. ROC-AUC values after 5-fold cross validation on the training dataset. Mean and standard deviation of ROC-AUC values are reported

Finally, both models were trained on the entire internal dataset and evaluated on the external dataset. It can be observed in Table 3 that for the breast cancer dataset, the results on the external dataset closely matched those in the internal dataset, with only a minor reduction in the ROC-AUC to 0.9892 for ClinicalBioBERT and 0.9883 for BERT-base. Using a threshold cutoff of 0.01 and BERT-base, we can estimate that for a randomly selected sample of 1,000 notes, only 54 would be flagged as positive and would require manual review. This is a 94.6% reduction in the volume of notes requiring manual review, at the expense of only 3 missed recurrences. Furthermore, on the colorectal cancer dataset, it can be observed that ClinicalBioBERT had a small edge as it achieved a ROC-AUC of 0.9810. Using a threshold cutoff of 0.01 and ClinicalBioBERT, we estimate that for a randomly selected sample of 1,000 notes, only 52 would be flagged as positive, requiring manual review, which is a 94.8% reduction. This would be at the expense of only 7 missed recurrences.

			Cut-off: 0.01		Cut-off: 0.5	
	ROC-AUC	Scl Br	Sn (PPV)	Sp (NPV)	Sn (PPV)	Sp (NPV)
Breast Cancer						
ClinicalBioBERT	.9892	.419	.926 (.592)	.981 (.998)	.863 (.659)	.987 (.996)
BERT-base	.9883	.377	.929 (.629)	.985(.996)	.863 (.639)	.986 (.996)
Colorectal Cancer						
ClinicalBioBERT	.9810	.251	.806 (.544)	.976 (.991)	.619 (.616)	.986 (.986)
BERT-base	.9694	.219	.751 (.531)	.976 (.991)	.577 (.612)	.987 (.985)

Table 3. Result values at multiple cutoffs on the breast cancer and colorectal cancer datasets. Scl Br: Scaled Brier Score. Sn: sensitivity. Sp: specificity. PPV: positive predictive value. NPV: negative predictive value

These results demonstrate that pre-trained transformer models can perform exceptionally well on detecting cancer recurrence from electronic medical record notes. These findings exceed previously reported results on cancer recurrence detection using classical machine learning methods [5], and earlier neural network architectures [2]. Results between BERT-base and ClinicalBioBERT showed only minor differences. In the internal datasets where cross-validation could be performed, the results were only a standard deviation away from each other. This suggests that the differences may not be statistically significant. The significance of this may lie in the fact that the dataset is of sufficient size for BERT to learn vocabulary associations with cancer recurrence and thus may take advantage of medical text pretraining less important.

A major advantage of the BERT approach is the model's ability to extract useful information without the need for expert knowledge. Earlier approaches using more traditional machine learning techniques required in-domain vocabulary compiled by oncologists [5], which is a time-consuming process. Moreover, since vocabulary and abbreviations may be local and institutional, BERT's approach eliminates the need for each institution to create its own vocabulary. These results pave the way for the use of this software in clinical work. Depending on clinical needs, a threshold cutoff can be selected to match the requirements. For example, to create automated tracking of the incidence of cancer recurrence, a lower threshold cutoff, such as 0.01, could be selected to achieve high sensitivity and specificity. Alternatively, to create a screening tool for patients with recurrence, generating curated lists that a human expert can then review, a higher threshold value closer to 0.5 would be selected to maximize specificity and minimize the likelihood of false negatives.

However, the main limitations of this study are that all data are obtained from the same institution, which may affect the generalizability of the trained model at other institutions. Moreover, it is possible that some cancer patients had notes spanning the periods of both the training and validation datasets, thereby causing similarities in some of their notes. Finally, the datasets were skewed towards negative occurrences, as is common in many medical datasets, and this resulted in a model with excellent specificity but relatively lower sensitivity. In conclusion, transformer-based models, such as BERT and its variants, can achieve excellent sensitivity and specificity on the task of detecting cancer recurrence from medical record notes. These models do not require rules or specialized knowledge from domain experts, yet nevertheless outperform earlier machine learning methods.

References

[1] Alsentzer E, Murphy J, Boag W, Weng W-H, Jin D, Naumann T, et al. Publicly Available Clinical BERT Embeddings. In: Proceedings of the 2nd Clinical Natural Language Processing Workshop [Internet]. Minneapolis, Minnesota, USA: Association for Computational Linguistics; 2019. p. 72–8. Available from: https://www.aclweb.org/anthology/W19-1909

[2] Banerjee I, Bozkurt S, Caswell-Jin JL, Kurian AW, Rubin DL. Natural Language Processing Approaches to Detect the Timeline of Metastatic Recurrence of Breast Cancer. JCO Clin Cancer Inform. 2019 Dec;(3):1–12.

[3] Bodenreider O. The Unified Medical Language System (UMLS): integrating biomedical terminology. Nucleic Acids Res. 2004 Jan 1;32(90001):267D – 270.

[4] Brenner DR, Weir HK, Demers AA, Ellison LF, Louzado C, Shaw A, et al. Projected estimates of cancer in Canada in 2020. Can Med Assoc J. 2020 Mar 2;192(9):E199–205.

[5] Carrell DS, Halgrim S, Tran D-T, Buist DSM, Chubak J, Chapman WW, et al. Using Natural Language Processing to Improve Efficiency of Manual Chart Abstraction in Research: The Case of Breast Cancer Recurrence. Am J Epidemiol. 2014 Mar 15;179(6):749–58.

[6] Denny JC, Peterson JF, Choma NN, Xu H, Miller RA, Bastarache L, et al. Extracting timing and status descriptors for colonoscopy testing from electronic medical records. J Am Med Inform Assoc. 2010 Jul;17(4):383–8.

[7] Devlin J, Chang M-W, Lee K, Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. ArXiv181004805 Cs [Internet]. 2019 May 24 [cited 2020 Apr 16]; Available from: http://arxiv.org/abs/1810.04805

[8] Hassett MJ, Uno H, Cronin AM, Carroll NM, Hornbrook MC, Ritzwoller D. Detecting Lung and Colorectal Cancer Recurrence Using Structured Clinical/Administrative Data to Enable Outcomes Research and Population Health Management. Med Care. 2017 Dec;55(12):e88–98.

[9] Johnson AEW, Pollard TJ, Shen L, Lehman LH, Feng M, Ghassemi M, et al. MIMIC-III, a freely accessible critical care database. Sci Data. 2016 May 24;3(1):160035.

[10] Lee J, Yoon W, Kim S, Kim D, Kim S, So CH, et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics. 2019 Sep 10;btz682.

[11] Michalopoulos G, Wang Y, Kaka H, Chen H, Wong A. UmlsBERT: Clinical Domain Knowledge Augmentation of Contextual Embeddings Using the Unified Medical Language System Metathesaurus. NAACL 2021.

[12] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention Is All You Need. ArXiv170603762 Cs [Internet]. 2017 Dec 5 [cited 2021 May 6]; Available from: http://arxiv.org/abs/1706.03762

Challenges of Trustable AI and Added-Value on Health
B. Séroussi et al. (Eds.)

doi:10.3233/SHTI220404

Characterization of Type 2 Diabetes Using Counterfactuals and Explainable AI

Marta LENATTI [a,1], Alberto CARLEVARO[a,b], Karim KESHAVJEE[c], Aziz GUERGACHI[d,e], Alessia PAGLIALONGA[a] and Maurizio MONGELLI[a]
[a] *National Research Council of Italy (CNR), Institute of Electronics, Information Engineering and Telecommunications (IEIIT), Italy;*
[b] *University of Genoa, Department of Electrical, Electronics and Telecommunications Engineering and Naval Architecture (DITEN), Italy;*
[c] *University of Toronto, Institute of Health Policy, Management and Evaluation, Dalla Lana School of Public Health, Canada*
[d] *Ryerson University, Ted Rogers School of Management, Toronto, Canada*
[e] *York University, Department of Mathematics and Statistics, Toronto, Canada*

Abstract. Type 2 diabetes mellitus is a metabolic disorder of glucose management, whose prevalence is increasing inexorably worldwide. Adherence to therapies, along with a healthy lifestyle can help prevent the onset of disease. This preliminary study proposes the use of explainable artificial intelligence techniques with the aim of (i) characterizing diabetic patients through a set of easily interpretable rules and (ii) providing individualized recommendations for the prevention of the onset of the disease through the generation of counterfactual explanations, based on minimal variations of biomarkers routinely collected in primary care. The results of this preliminary study parallel findings from the literature as differences in biomarkers between patients with and without diabetes are observed for fasting blood sugar, body mass index, and high-density lipoprotein levels.

Keywords. Diabetes, Counterfactual Explanations, eXplainable AI

1. Introduction

An increasing number of adults are at risk of developing Type 2 Diabetes (T2D) worldwide, correlating with trends in population aging and sedentary lifestyles. In 2021, an estimated 61.4 million (9.2%) of adults in Europe were living with diabetes, 90% of whom had T2D [1]. Interventions aimed at reducing the risk of T2D and delay disease progression can help reduce its massive burden (European diabetes-related expenditure: US$ 189 billion). To support T2D prediction and prevention, several classification methods have been proposed (e.g., [2]). In the context of explainable artificial intelligence, which is gaining increasing attention in the medical field [3], counterfactual explanations (CEs) [4] are a local technique that searches for minimal distance variations in the input features space (e.g., biomarkers), that would lead to a change in the output class (e.g., diabetes diagnosis). Recently, different approaches for CEs generation have been proposed (e.g., deep learning based [5] or graph based [6]). For example,

[1] Corresponding Author: Marta Lenatti, CNR-IEIIT, c/o Politecnico di Milano, Piazza Leonardo da Vinci, 32, I-20133 Milan, Italy; Phone: +02 23999648; E-mail:marta.lenatti@ieiit.cnr.i

Dhurandhar et al [7] analyzed local explanations applied to brain images to understand why certain subjects were classified as autistic and others were not. To our knowledge, none of the previously proposed methodologies have assessed CE using a rule-based description of the output classes to provide personalized treatments.

The aim of this study was to introduce and characterize CEs as a method to develop individualized diabetes prevention recommendations. We investigate biomarkers, and change in biomarkers, that can help reduce the individual risk of T2D onset and progression.

2. Methods

The dataset used in this study includes 1857 records extracted from primary care electronic medical records (EMRs) from a multi-disease national database, the Canadian Primary Care Sentinel Surveillance Network (CPCSSN, http://cpcssn.ca/). Specifically, the dataset includes 428 subjects with T2D (class "T2D"; age: mean= 56, range=18-90; gender: 227 F, 201 M) and 1429 subjects without T2D (class "No T2D"; age: mean= 63, range=18-83; gender: 916 F, 513 M). The prevalence of T2D in our dataset is 23% and it reflects the prevalence of diabetes in older adults (around 20% in adults > 60 years) [1]. For each patient, seven biomarkers were considered: systolic blood pressure (sBP), body mass index (BMI), low-density lipoprotein (LDL), high-density lipoprotein (HDL), triglycerides (TG), fasting blood sugar (FBS), and age. A Two Class Support Vector Data Descriptor (TC-SVDD) [8] was trained on 80% of the dataset by applying stratification and a natively explainable method, the Logic Learning Machine (LLM) [9], was used to extract a rule-based description of the two classification regions, following the approach presented in [10]. Counterfactual explanations were generated from a set of 120 randomly sampled points within the SVDD region describing T2D subjects. The sample size was defined following preliminary analysis to ensure adequate coverage of the whole age range, including the tails of the distribution (i.e., <50 years and >75 years). The rationale in CE generation was to find the minimum perturbations that should be applied to each biomarker to allow a subject initially classified as diabetic by the TC-SVDD, to be classified as non-diabetic. Given a binary classification problem characterized by two classes S_1 and S_2 and an observation $\boldsymbol{x} \in S_1$, our goal is to determine the minimum variation $\Delta\boldsymbol{x}^*$, so that the point $\boldsymbol{x}^* = \boldsymbol{x} + \Delta\boldsymbol{x}^*$ belongs to the opposite class, by solving the following minimization problem:

$$\underset{\Delta\boldsymbol{x} \in \mathbb{R}^n}{min}\, d\big(\boldsymbol{x}, (\boldsymbol{x} + \Delta\boldsymbol{x})\big)\ subject\ to\ \|(\boldsymbol{x} + \Delta\boldsymbol{x}) - \boldsymbol{a}_2\|^2 \le R_2^2 \wedge \|(\boldsymbol{x} + \Delta\boldsymbol{x}) - \boldsymbol{a}_1\|^2 \ge R_1^2$$

where R_1^2, R_2^2, $\boldsymbol{a}_1$, $\boldsymbol{a}_2$ are the radii and the centers of the spheres for class S_1 and S_2, respectively, as defined by the TC-SVDD. The algorithms were implemented in MATLAB (version 2021a).

3. Results

The TC-SVDD achieved the following classification performance on the test set: accuracy = 0.77, sensitivity = 0.65, and specificity = 0.83. Table 1 shows the rules extracted by the LLM from the two regions identified by the TC-SVDD, and the rule

covering (i.e., the percentage of records in the output class that satisfy the rule). Consistent with the diagnostic criteria for T2D [1], the two rules that best characterize the two classes (i.e., the rules with higher covering), are based on the FBS value alone. In particular, high FBS values (i.e., higher than 6.89 mmol/L) are associated with class "T2D" (rule #1) whereas low FBS values (i.e., lower than 5.1 mmol/L) are associated with class "No T2D" (rule #5). The remaining records are covered by more complex rules, i.e., rules based on more than one condition.

Table 1. Rules characterizing the two SVDD regions and rule covering (Cov).

#	Class	Rule	Cov%
1	T2D	FBS>6.89	52.70
2	T2D	FBS>4.91 ∧ 48< age ≤80 ∧ 98< sBP ≤163 ∧ 21< BMI ≤48 ∧ LDL≤1.88	19.90
3	T2D	6.33 < FBS ≤ 6.91 ∧ age > 37 ∧ 100 < sBP ≤ 145	16.55
4	T2D	6.04 < FBS ≤ 6.31 ∧ age ≤ 76 ∧ sBP ≤ 152 ∧ 22 < BMI ≤ 48 ∧ TG > 0.17	8.16
5	NoT2D	FBS ≤ 5.1	47.46
6	NoT2D	FBS ≤ 7.23 ∧ sBP ≤ 111	18.22
7	NoT2D	FBS ≤ 6.68 ∧ age > 50 ∧ BMI > 24 ∧ LDL > 3.61	15.57
8	NoT2D	FBS ≤ 7.87 ∧ 147 < sBP ≤ 182 ∧ LDL > 1.09	15.51

An example of counterfactual (class "No T2D") generated from a record of a T2D patient under the constraint of minimal distance is shown below:

- *T2D patient:* Age = 73 years; sBP = 113 mmHg; BMI = 29.4 kg/m2; LDL = 2.3 mmol/L; HDL = 1.82 mmol/L; TG = 0.83 mmol/L; FBS = 6.9 mmol/L
- *Counterfactual*: Age = 73 years; sBP = 113 mmHg; BMI = 27.9 kg/m2; LDL = 2.38 mmol/L; HDL = 3 mmol/L; TG = 1.07 mmol/L; FBS = 4.89 mmol/L

Small, yet meaningful changes can be found between the observation as the diabetic patient is associated with higher FBS and BMI, stable sBP and lower HDL, LDL, and TG, with respect to the counterfactual, allowing for the two observations to be associated with different classes, hence to be described by different rules, as shown in Table 1.

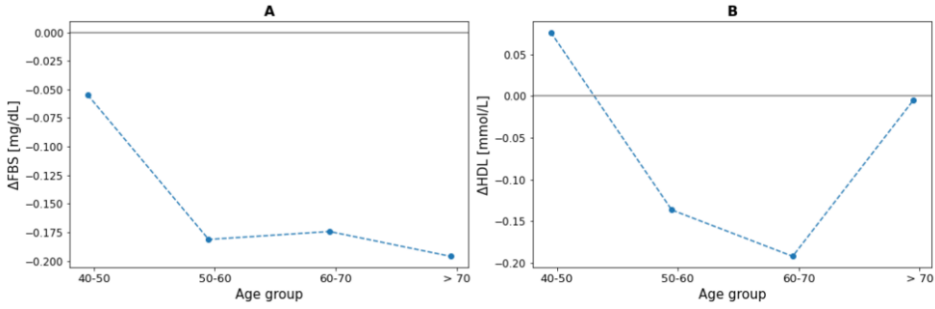


Figure 1. Average change in FBS (panel A) and HDL (panel B) associated with a variation in the output class from class "T2D" to class "No T2D" as a function of the age group.

The average change observed for each biomarker, computed as the difference between the value in the counterfactual (i.e., class "No T2D") and the value in the actual T2D patient is shown in Figure 1 as a function of age (i.e., 40-50 years, 50-60 years, 60-70 years, and 70 or more years). Figure 1(A) shows that the average change in FBS is negative, indicating an improvement in FBS from a diabetic subject to a non-diabetic one. The change in FBS tends to increase with age, ranging, on average, from -0.2 to -1.2 mmol/L. Vice versa, Figure 1(B) shows that the average variation for HDL is positive,

as non-diabetic subjects are characterized by higher values (i.e., better values) with respect to diabetic ones. Regarding HDL, the variation is quite stable for different age groups, around 0.20-0.30 mg/dL.
The average variation observed across the 120 generated counterfactuals is: ΔAge = 0.01 years; ΔsBP = −0.17 mmHg; ΔBMI = −0.14 kg/m2; ΔLDL = 0.53 mmol/L; ΔHDL = 0.26 mmol/L; ΔTG = 0.12 mmol/L; ΔFBS = −0.88 mmol/L.

4. Discussion

The extraction of rules from the two classification regions identified by the TC-SVDD allows us to obtain a comprehensive, readable view of the observed pathology, that can be easily interpreted and validated by clinicians, even if they have no prior knowledge in the field of artificial intelligence. Counterfactual explanations allow us to move from a global to a local approach, tailored on an individual basis.

Although the counterfactual generation process searches, by its definition, for minimal variations that determine a change in output class, the results obtained allow us to appreciate differences in biomarkers between the two classes, in line with the available knowledge. Indeed, high FBS is a peculiar characteristic of diabetic subjects, as T2D is a metabolic disorder of glucose regulation resulting in hyperglycemia. A trend towards higher BMI in diabetic patients is consistent with the known link between obesity and T2D. For this reason, maintaining a healthy lifestyle characterized by moderate exercise can help in the prevention of diabetes. Moreover, low HDL concentrations have been associated with higher risk of developing T2D [11], possibly representing one of the risk factors for cardiovascular disease. In contrast, a counterintuitive result is related to the presence of slightly higher (i.e., worse) LDL and TG values in nondiabetic patients, compared with diabetic ones. This result can be explained by the fact that diabetic and pre-diabetic subjects are very often prescribed LDL-lowering drugs like statins [12] to control hypercholesterolemia. Further research on larger data should characterize different stratifications, such as by gender, as the variations obtained in biomarkers might be different for male or female subjects. In addition, causal relationships among biomarkers should also be assessed, for example using causal inference approaches.

The proposed method is characterized by a degree of uncertainty in the classification, because counterfactuals are sought in the proximity of the decision boundary, defined by the SVDD, thus only decreasing the risk of developing diabetes by a small amount (the minimal amount needed for changing the output class). A safer definition of subjects with low probabilities of developing diabetes, would be the search of counterfactuals from “reliable SVDD regions”, i.e., by defining regions with a reduced false negative rate, as in [13]. The reliable approach is based on the relaxation of the definition of counterfactuals generated at minimum distance, as they will be searched within spherical regions of smaller size, enabling a more precise identification of the opposite class. Specifically, the TC-SVDD will define a region surrounding non-diabetic subjects, minimizing the presence of diabetic subjects within that region, according to the imposed false negative rate. In this way, personalized prevention strategies could be devised through more effective changes in biomarkers, so that the risk of developing the disease can be substantially reduced. Also, it should be noted that characterization of diabetic patients is based on a combination of controllable characteristics (biomarkers that can be manipulated through therapies or lifestyle changes) and non-controllable characteristics (e.g. features that are not manipulable by definition such as age, genetic factors, and

family history). Therefore, the search for realistic counterfactuals should be performed by perturbing only controllable variables and keeping non-controllable variables fixed.

5. Conclusions

This preliminary study aims to demonstrate the efficacy of a new method for CE generation based on SVDD on biomarker variation and diabetes diagnosis. Specifically, the search for minimal variations in biomarkers, able to change the classification of the subject from diabetic to non-diabetic allows us to highlight a trend (decrease or increase) in biomarkers, consistent with current knowledge. Further research is needed to validate the proposed method in different datasets (e.g., with varying size and number of output classes) and for different chronic diseases (e.g., COPD). This approach can be extended in the future to a broader set of biomarkers including lifestyle indicators (e.g., diet, alcohol consumption, smoke), comorbidities and medications and to a broader definition of counterfactual obtained perturbing controllable features only, at progressively increasing distance (i.e., progressively decreasing risk of developing diabetes), with the ultimate goal of personalizing prevention strategies on an individual basis.

Acknowledgments: The work was supported by Fondazione Compagnia di San Paolo, scientific research call 2019 (Bando 2019–2020 per progetti di ricerca scientifica presentati da enti genovesi): project “Advances in pneumology via ICT and data analytics” (PNEULYTICS)

References

[1] International Diabetes Federation (2021). IDF Diabetes Atlas 10th edition.
[2] Silva KD, Lee WK, Forbes A, Demmer RT, Barton C, Enticott J. Use and performance of machine learning models for type 2 diabetes prediction in community settings: A systematic review and meta-analysis. International Journal of Medical Informatics. Elsevier BV; 2020. p. 104268.
[3] Hakkoum H, Abnane I, Idri A. Interpretability in the medical field: A systematic mapping and review study. Applied Soft Computing; 2022. 117. p. 108391
[4] Wachter S, Mittelstadt BDM, and Russell C. Counterfactual Explanations without Opening the Black Box: Automated Decisions and the GDPR. Harvard Journal of Law and Technology. 2018;31(2):841–87.
[5] Nemirovsky D, Thiebaut N, Xu Y, Gupta A, “Providing actionable feedback in hiring marketplaces using generative adversarial networks,” WSDM ’21. Association for Computing Machinery, 2021.
[6] Poyiadzi R, Sokol K, Santos-Rodríguez R, Bie TD, Flach PA, “Face: Feasible and actionable counterfactual explanations,” AAAI/ACM Conference on AI, Ethics, and Society, 2020.
[7] Dhurandhar A, Chen PY, Luss R, Tu CC, Ting P, Shanmugam K, Das P. Explanations based on the missing: towards contrastive explanations with pertinent negatives. In Proceedings of the 32nd International Conference on Neural Information Processing Systems (NIPS'18), 2018.
[8] Huang G, Chen H, Zhou Z, Yin F, Guo K. Two-class support vector data description. Pattern Recognition. Elsevier BV; 2011. p. 320–329.
[9] Muselli M, Ferrari E. Coupling Logical Analysis of Data and Shadow Clustering for Partially Defined Positive Boolean Function Reconstruction. IEEE Trans. Knowl. Data Eng. IEEE; 2011. p. 37–50.
[10] Carlevaro A, Mongelli M. Reliable AI Through SVDD and Rule Extraction. Lecture Notes in Computer Science. Springer International Publishing; 2021. p. 153–171.
[11] Wilson PWF. Prediction of Incident Diabetes Mellitus in Middle-aged Adults. Arch Intern Med. American Medical Association (AMA); 2007. p. 1068.
[12] Feingold KR. Cholesterol Lowering Drugs. 2021 Mar 30. In: Feingold KR, Anawalt B, Boyce A,Chrousos G, de Herder WW, Dhatariya K, et al, editors. Endotext.

[13] Carlevaro A, Mongelli M. A New SVDD Approach to Reliable and eXplainable AI. IEEE Intell. Syst.IEEE; 2021. p. 1–1.

Challenges of Trustable AI and Added-Value on Health
B. Séroussi et al. (Eds.)

doi:10.3233/SHTI220405

Utilizing a Non-Motor Symptoms Questionnaire and Machine Learning to Differentiate Movement Disorders

Alexander BRENNER[a], Lucas PLAGWITZ[a], Michael FUJARSKI[a], Tobias WARNECKE[b] and Julian VARGHESE[a]
[a] *Institute of Medical Informatics, University of Münster, Münster, Germany*
[b] *Department of Neurology and Neurorehabilitation, Klinikum Osnabrück – Academic teaching hospital of the University of Münster, Osnabrück, Germany*

Abstract. Parkinson's disease (PD) is a common neurodegenerative disorder that severely impacts quality of life as the condition progresses. Early diagnosis and treatment is important to reduce burden and costs. Here, we evaluate the diagnostic potential of the Non-Motor symptoms (NMS) questionnaire by the International Parkinson and Movement Disorder Society based on patient-completed answers from a large single-center prospective study. In this study data from 489 study participants consisting of a PD group, a healthy control (HC) group and patients with differential diagnosis (DD) have been recorded with a smartphone-based system. Evaluation of the study data has shown a significant difference in NMS between the representative groups. Cross-validation of Machine Learning based classification achieves balanced accuracy scores of 88.7% in PD vs. HC, 72.1% in PD vs. DD and 82.6% when discriminating between all movement disorders (PD + DD) and the HC group. The results indicate potentially high feature importance of a simple self-administered questionnaire that could support early diagnosis.

Keywords. Parkinson's Disease, movement disorders, artificial intelligence

1. Introduction

Worldwide prevalence of Parkinson's disease (PD) has more than doubled over the last two decades [1]. It affects movement with symptoms such as tremor, slowness and rigidity. However, non-motor symptoms are often reported as early indicators of the disease that can severely impact patient's quality of life [2]. Currently, PD is diagnosed primarily on the basis of clinical examination and nuclear imaging, along with other movement disorders. Approaching the need for comprehensive evaluation of mHealth-based biomarkers, the Smart Device System (SDS) has been implemented and applied in a three-year prospective study [3]. From 2018 to 2021, a total of 489 participants, including a broad range of different PD progress states according to Hoehn and Yahr [4] have been recorded. Each session comprises a series of sensor measures of movements, speech recordings, tablet-based tasks and smartphone-based questionnaires. With the recent conclusion of the aforementioned study, here, we focus on diagnostic potential of the Non-Motor symptoms (NMS) questionnaire by the International Parkinson and Movement Disorder Society as preliminary findings of the study.

While the NMS are routinely used to document early symptoms in PD, we elaborate on the diagnostic potential by including PD patients, healthy controls and other movement disorders as well. Our research question is whether we can reliably detect and discriminate PD from healthy controls and other movement disorders using the smartphone-based questionnaire data from our study dataset. We therefore evaluated Machine Learning models for three different classification tasks and pointed out possible important feature groups in the patient data. With the analysis of the questionnaire, which is a simple patient reported outcome, we have performed an important step in the evaluation of our study with the objective of finding digital biomarkers to aid PD diagnosis or screening in the future.

2. Methods

The study has been registered (ClinicalTrials.gov ID: NCT03638479) and approved by the ethical board of the University of Münster and the physician's chamber of Westphalia-Lippe (Reference number: 2018-328-f-S). Our study data consist of PD patients, a HC group and patients with DD. All diagnoses were confirmed by neurologists and reviewed by one senior movement disorder expert. The participant sample is summarized in Table 1. The first step of examination provides short questionnaires about medical history and non-motor symptoms (NMS) via a smartphone app. The medical history questionnaire holds information about age, height, weight, kinship with PD and effect of alcohol on tremor (further details provided in Varghese et al. [3]). The self-completed symptoms questionnaire is based on Chaudhuri et al. [5] and composed of 30 yes/no items. It represents different areas of possible PD symptoms.

First, we analyzed the NMS score, i.e., the sum of yes responses per participant, to investigate simple group separation. For comparability, we calculated statistical significance between the representative groups. In order to investigate the supportive potential in a clinical examination scenario, we trained and evaluated different Machine Learning algorithms via cross-validation. Given the questionnaire data, three relevant classification tasks were approached: (1) PD vs. HC, (2) Movement disorders (PD + DD) vs. HC and (3) PD vs. DD. To obtain stable results, a 5-fold cross-validation was randomized in 5 replicates, so that an average score over 25 test sets was produced. Three Machine Learning estimators were evaluated: (1) a multi-layer perceptron neural network (MLP), (2) a support-vector machine (SVM) and (3) CatBoost [6], which is based on gradient-boosted decision trees. All classifiers were trained with balanced weights, as class distributions are unequal in our dataset. Similarly, comparison of classification performance is based on balanced accuracy, precision and recall. Further details on hyper-parameter settings can be found in the supplement [7]. Based on the overall best performing classifier, feature importance was analyzed via permutation importance analysis using Scikit-Learn (0.24.2) [8]. Since NMS questions are grouped by relevant domains and thus certain questions are linked together, we modified the algorithm to consider and permute features group-wise. The selected feature-groups are based on the originally intended domains [5], details are provided in the supplements [7].

Table 1. Participant sample. Height in centimeters, weight in kilograms. Values correspond to mean (SD).

Disease Class	#Samples	Age	Height	Weight	NMS score
Parkinson's Disease (PD)	276	65.5 (9.6)	170.9 (15.1)	78.7 (16.8)	9.93 (5.25)
Differential diagnosis (DD)	124	60.2 (13.4)	175.6 (9.8)	83.6 (18.2)	7.42 (4.92)
Healthy control (HC)	89	60.2 (14.8)	173.5 (8.6)	79.0 (17.7)	2.36 (2.69)

3. Results

Table 1 shows the NMS score across the groups. On average NMS scores are lowest in the healthy group (mean 2.36, median 2), followed by the DD group (mean 7.42, median 7) and highest for PD patients (mean 9.93, median 9). In pairwise comparison the groups significantly differ from each other in NMS scores (Mann-Whitney, Kruskal-Wallis and t-test, $P < 0.0001$). Percentages of yes-answers per question are presented in the supplements [7].

To analyze the potential of the questionnaire data in supporting diagnosis, the previously described Machine Learning models were applied to the study data. Classifiers were cross-validated in different settings: (1) using only NMS score, (2) using all NMS items and (3) using the complete questionnaire (except medication). Complete results can be found in the supplements [7]. The overall best performing classifier scores are summarized in Table 2.

Table 2. Performance for the tasks (1) PD vs. HC, (2) Movement disorders (PD + DD) vs. HC and (3) PD vs. DD. All results are based on CatBoost classifier. Values correspond to mean (SD). *Balanced Accuracy.

Task	Input	Accuracy	B. Accuracy*	Precision	Recall
PD vs. HC	NMS score	0.817 (0.036)	0.849 (0.047)	0.965 (0.032)	0.787 (0.039)
	NMS quest.	0.886 (0.049)	0.887 (0.045)	0.961 (0.025)	0.885 (0.048)
	Full quest.	0.888 (0.041)	0.873 (0.047)	0.947 (0.025)	0.902 (0.046)
PD + DD vs. HC	NMS score	0.755 (0.040)	0.816 (0.039)	0.974 (0.019)	0.720 (0.050)
	NMS quest.	0.831 (0.036)	0.810 (0.042)	0.945 (0.018)	0.843 (0.044)
	Full quest.	0.860 (0.032)	0.826 (0.057)	0.946 (0.024)	0.879 (0.032)
PD vs. DD	NMS score	0.623 (0.043)	0.610 (0.037)	0.772 (0.029)	0.646 (0.075)
	NMS quest.	0.740 (0.044)	0.702 (0.054)	0.820 (0.039)	0.800 (0.052)
	Full quest.	0.761 (0.040)	0.721 (0.049)	0.828 (0.034)	0.826 (0.045)

Using the complete questionnaire information, grouped feature importance analysis was performed for task 1, PD vs. HC. The computed information gain per feature-group is depicted in Figure 1.

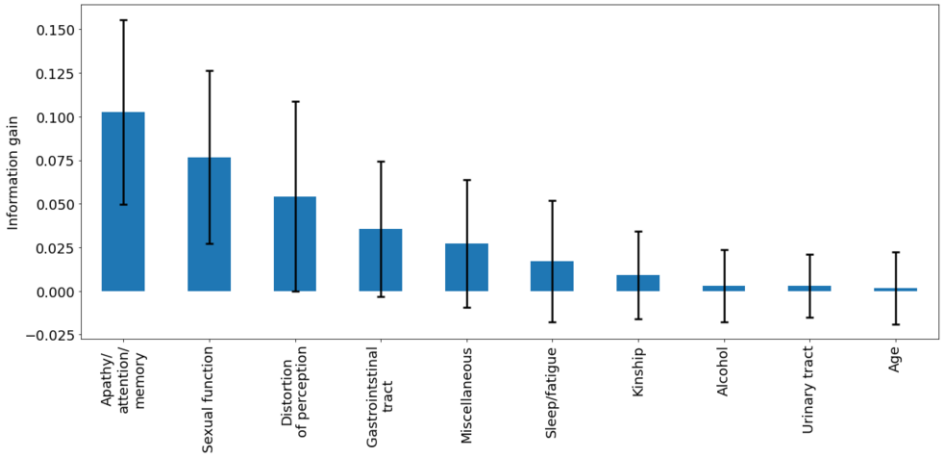


Figure 1. Grouped feature importance analysis of the NMS questionnaire. Bar height corresponds to mean information gain (± SD). Top 10 feature groups are displayed.

4. Discussion

We have evaluated the diagnostic accuracy of self-completed questionnaire data from our recently closed single-center prospective study for the patient groups PD, DD and HC. By testing different classification models and subsets of questions, we explored a robust set of features for disease prediction. For PD vs. HC the highest balanced accuracy score of 88.7% is achieved using NMS questions without additional information. Full questionnaire data resulted in a score of 87.3%. Given that additional data does not improve classification performance, we believe that the NMS questionnaire itself provides highest information gain in identifying PD patients. Our observations are similar in the task PD + DD vs. HC. Here, balanced classification accuracy for the NMS questionnaire, 81.0%, and the full questionnaire, 82.6%, are almost equal. It also shows that discrimination between healthy samples and PD is more accurate than discrimination between the healthy group and all movement disorders. This is to be expected as the questionnaire has been design to cover PD specific NMS. While the most challenging task - PD vs. DD - achieves lowest balanced accuracy overall, the highest score of 72.1% using the full questionnaire still discriminates both groups reasonably well. Here, the score for just the NMS questionnaire is slightly lower with 70.2%. It is noteworthy that even the aggregated NMS score already can reasonably divide the target groups, especially PD vs. HC with 84.9% balanced accuracy. Nonetheless, how PD affects a person can highly vary from patient to patient. Finding specific symptoms and subgroups can influence diagnosis and consequently treatment.

In our feature importance analysis for PD vs. HC, the top three groups in terms of information gain are apathy/attention/memory, sexual function and distortion of perception. This observation partially confirms the findings of previous studies [5,9]. In contrast, comparing to a comprehensive Italian study with a structured interview and an extended NMS questionnaire, fatigue and anxiety have shown to be more prevalent in PD patients [10]. Differences can arise from various causes, i.e. differences in answering (self-completed or structured), differences between study groups, while also some NMS generally become more common with higher age [2].

A limitation of our study is that it is conducted at a single site and thus bias due to specific site population, examination setting or target diagnoses can not be ruled out. Still, we believe the study data provides a suitable representation, as participants have been recruited from a large tertiary care center and include various types and stages of PD and other movement disorders. An important issue in Machine Learning remains the lack of interpretability and transparency when using complex models or feature computations. As a result, there is often mistrust in using models as decision support, even if they are able to make accurate predictions when measured against the gold standard. In our case, we worked with structured tabular data so that input features remained interpretable. Further, the feature importance analysis gave us insights about what questions were relevant for the decision process. We have demonstrated that the NMS yes/no answers were sufficient to classify PD from healthy controls with 88.7% accuracy. Overall it can be concluded, that a simple symptoms questionnaire could be of high importance for early patient screening. With the inclusion of additional sensor-based modalities from the study we will further continue the analysis.

5. Conclusion

A simple symptoms questionnaire and supervised Machine Learning provide high classification accuracy in the domain of Movement Disorders, in particular Parkinson's disease. While this questionnaire is known in practice to document non motor symptoms of Parkinson's disease we show its diagnostic potential against healthy participants and even differential diagnoses via Machine Learning. It could pave the way for early patient screening utilizing mHealth technologies. Continuing the work, we will further analyze the study data including further sensor-based modalities by means of dimensionality reduction and clustering techniques in order to find possible subgroups of the disease.

References

[1] Rocca WA. The burden of Parkinson's disease: a worldwide perspective. The Lancet Neurology. 2018;17(11):928-9.
[2] Pfeiffer RF. Non-motor symptoms in Parkinson's disease. Parkinsonism & related disorders. 2016;22:S119-22.
[3] Varghese J, Niewöhner S, Soto-Rey I, Schipmann-Miletić S, Warneke N, Warnecke T, et al. A smart device system to identify new phenotypical characteristics in movement disorders. Frontiers in neurology. 2019;10:48.
[4] Bhidayasiri R, Tarsy D. Parkinson's disease: Hoehn and Yahr scale. In: Movement Disorders: A Video Atlas. Springer; 2012. p. 4-5.
[5] Chaudhuri KR, Martinez-Martin P, Schapira AH, Stocchi F, Sethi K, Odin P, et al. International multicenter pilot study of the first comprehensive self-completed nonmotor symptoms questionnaire for Parkinson's disease: the NMSQuest study. Movement disorders: official journal of the Movement Disorder Society. 2006;21(7):916-23.
[6] Dorogush AV, Ershov V, Gulin A. CatBoost: gradient boosting with categorical features support. arXiv preprint arXiv:181011363. 2018.
[7] Brenner A, Plagwitz L, Fujarski M, Warnecke T, Varghese J. Online supplement SDS MIE 2022; 2022. [cited 2022-01-23]. Available from: https://uni-muenster.sciebo.de/s/XIGIMyI1gLMsEbY.
[8] Buitinck L, Louppe G, Blondel M, Pedregosa F, Mueller A, Grisel O, et al. API design for machine learning software: experiences from the scikit-learn project. In: ECML PKDD Workshop: Languages for Data Mining and Machine Learning; 2013. p. 108-22.
[9] Khoo TK, Yarnall AJ, Duncan GW, Coleman S, O'Brien JT, Brooks DJ, et al. The spectrum of nonmotor symptoms in early Parkinson disease. Neurology. 2013;80(3):276-81.
[10] Barone P, Antonini A, Colosimo C, Marconi R, Morgante L, Avarello TP, et al. The PRIAMO study: a multicenter assessment of nonmotor symptoms and their impact on quality of life in Parkinson's disease. Movement disorders: official journal of the Movement Disorder Society. 2009;24(11):1641-9.

Challenges of Trustable AI and Added-Value on Health
B. Séroussi et al. (Eds.)

doi:10.3233/SHTI220406

Supporting AI-Explainability by Analyzing Feature Subsets in a Machine Learning Model

Lucas PLAGWITZ[a,1] Alexander BRENNER[a], Michael FUJARSKI[a], and Julian VARGHESE[a]
[a]*Institute of Medical Informatics, University of Münster, Germany*

Abstract. Machine learning algorithms become increasingly prevalent in the field of medicine, as they offer the ability to recognize patterns in complex medical data. Especially in this sensitive area, the active usage of a mostly black box is a controversial topic. We aim to highlight how an aggregated and systematic feature analysis of such models can be beneficial in the medical context. For this reason, we introduce a grouped version of the permutation importance analysis for evaluating the influence of entire feature subsets in a machine learning model. In this way, expert-defined subgroups can be evaluated in the decision-making process. Based on these results, new hypotheses can be formulated and examined.

Keywords. explainable AI, permutation importance, grouped variable analysis

1. Introduction

Whether in clinical research or risk factor calculations, machine learning algorithms can be found in many medical fields. This results in important challenges for the usage of these algorithms - especially regarding the ethical background - like regulatory aspects, interpretability, or interoperability [1]. For this work, we present a particular perspective on interpretability of feature contributions. Namely, this should not be seen as just a means of application, but rather a way to better understand the underlying problem. Our focus is on the combination of humanly understandable feature subgroups and their respective importance scores.

Feature importance analysis is a major component in the examination and interpretation of machine learning models. The question arises on which basis a problem was solved, e.g., in classification, why the decision for a class was made the way it was. The methodology behind such an analysis can vary widely depending on the problem and model. We focus in the following on the permutation importance analysis [2]. Here, the variation of predictive accuracy is analyzed based on the test set. The feature under investigation is permuted and the prediction is compared with the unmodified one. Therefore, this method is not dependent on a specific implementation of the prediction process and thus can be applied to any feature-based supervised machine learning model.

[1] Corresponding Author, Lucas Plagwitz, Institute of Medical Informatics, University of Münster, Münster, Germany; E-mail: lucas.plagwitz@uni-muenster.de.

This allows a wide range of application of this method, even for more complicated architectures such as deep neural networks.

In this paper we are concerned with a particular adaptation of the permutation importance analysis. Many data sets consist of features that can be grouped to clinically meaningful subsets. An arrangement into such subsets does not have to be unambiguous at all but still relevant for the domain. We present a methodology to determine the information gain for each of these groups. In particular, for tree-based classifications, analyses on grouped feature importance and terms like group permutation importance measures have been described previously [3]. Their research supports the theoretical aspect of our application. As a complement to this, we intend to focus more on interpretation - by using expert-defined feature subsets - than optimization of the model. To this end, we present an explicit algorithm that uses cross-validation to provide a fast, stable, and meaningful analysis based on the test performance. In doing so, we shed some light on a black box prediction. In addition to the method itself, we focus on an application in neuroimaging where we evaluate the impact of a brain region depending on the target variable.

2. Methods

In the context of grouped characteristics, it is not advisable to simply add up individual feature importance scores. The reasons depend on the method of calculating scores. For example, in the univariate permutation importance analysis, it is assumed that all other features are known at the time of permutation. This does not provide adequate insight into the influence of a specific feature group. Other feature importance methods, such as a tree-based calculation, tend to overfit, often boosting features that contribute hardly any information to the problem [4]. Thus, a sum of these feature weights overestimates the group with substantially more features, while the mean or median discriminates against this group.

For this reason, we adjust the calculation as follows: First, a machine learning model is trained based on the training set and a feature subset is specified. Then, a predefined number $T \in \mathbb{N}$ of test set permutations are created. Here, it is important that only the features within the subset are permuted. Predictions are then created based on these modified test sets. By matching the correct labels, the change in performance is estimated (averaged over all permutations). This process is illustrated in Figure 1 and is repeated for every feature subset, starting with the existing trained model.

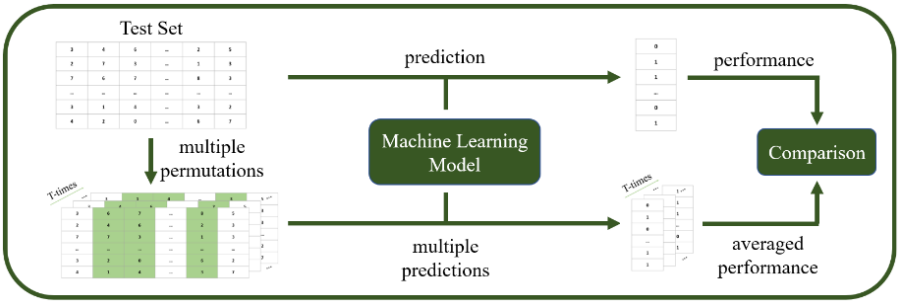


Figure 1. The permutation process of a specific feature subset is shown. The columns that are affected by the permutation are marked in green, the remaining columns remain unchanged.

Algorithm 1 Grouped Permutation Importance

Input: untrained model M; data X; targets y; k different feature subsets g_i; repeated permutation count T; cross-validation object CV, scoring-metric $m(\cdot,\cdot)$
Output: grouped permutation importance GPI $\in \mathbb{R}^k$

GPI $= 0^{(k)}$
for $(idx_{train}, idx_{test}) \in \text{CV}.split(X)$ **do**
  $\text{M}.fit(X[idx_{train}], y[idx_{train}])$
  $y_0 = \text{M}.predict(X[idx_{test}])$
  for $i = 1, \ldots, k$ **do**
    for $t = 1, \ldots, T$ **do**
      create permutation matrix P_t
      Set $y_{g_i}^{(t)}$ to the prediction from the column-wise permutation P_t in columns g_i of $X[idx_{test}]$.
      $\text{GPI}[i] = \text{GPI}[i] + (m(y[idx_{test}], y_0) - m(y[idx_{test}], y_{g_i}^{(t)}))/(|cv\_splits| \cdot T)$
    end for
  end for
end for

To stabilize the validity, this process is repeated over several cross-validation folds. The final value is then composed of the mean between these folds. Algorithm 1 summarizes the whole method. The corresponding code is published as open source[2]. The grouped permutation importance (GPI) describes the direct influence compared to the respective overall performance. Thus, it represents the absolute information gain per group. However, these analyses are particularly interesting in comparison with the same data set but different target variables. Since the maximum performance of a model depends strongly on the underlying target, a relative value should be preferred to an absolute one in these cases. For this purpose, the GPI is set in relation to the unpermuted performance to be able to analyze a relative information gain.

3. Experiment: Brain region impact depending on the target variable

The explanation of machine learning algorithms in the field of neuroimaging is a widespread analysis. One is interested in specific brain markers to gain a more detailed insight into the functionality of the brain. In the following, we see that depending on the target variable, expert-defined brain areas are incorporated as feature groups into the machine learning model on very different weights.

Based on a predefined atlas, the measured voxels were clustered and averaged into feature subgroups that we can evaluate by applying Algorithm 1. For the simulation, we use the public data set OASIS [5]. In addition to the T1-weighted magnetic resonance imaging (MRI) scans of 403 subjects, this data set provides three target variables: age, biological sex, and the clinical dementia rating (CDR), which is used for the diagnosis

[2] The code is available on GitHub and the installation is possible via pip: pip install git+https://github.com/lucasplagwitz/grouped_permutation_importance

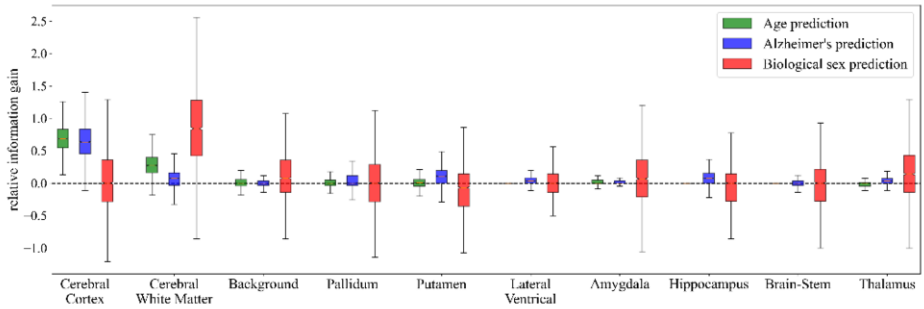


Figure 2. Brain region influence depending on the target variable. The displayed relative scores are calculated based on the balanced accuracy score. The following performances depending on the classification task were obtained: age 87%, Alzheimer's 82%, and biological sex prediction 72 %.

of Alzheimer's disease. For simplicity, we change the target variables to binary values. We only distinguish between CDR zero or greater than zero for Alzheimer's prediction and age more or less than 50 for an age-based classification task. By applying the Harvard-Oxford 2mm Subcortical Atlas, we obtain several volumes for different brain areas as seen in Table 1.

Cerebral Cortex	Cerebral White Matter	Putamen	Pallidum	Background	Lateral Ventrical
4	13	2	2	2	1

Brain-Stem	Amygdala	Hippocampus	Thalamus
1	2	2	2

Table 1. Havard-Oxford (Subcortical) Atlas – volumes per region.

All atlas calculations are based on the Python package nilearn [6]. We then determine the GPI of each brain region given a balanced support vector classifier, 5-fold cross-validation (randomized in 10 replicates), and a permutation value T = 100 by using the Python package scikit-learn [7], which is visualized in Figure 2. It is noticeable that the information about age and Alzheimer's disease is mainly located in the cerebral cortex. However, the cerebral white matter contains much more information about age. In contrast, the prediction of biological sex, this is based mainly in the white matter and hardly in the cerebral cortex. Moreover, the influence of the putamen as well as the hippocampus should be emphasized. While these regions provide no additional benefit for the prediction of age or biological sex, a positive impact on predicting Alzheimer's disease can be determined. Findings such as these could lead to a better differentiation of Alzheimer's disease from the normal aging process.

4. Discussion

We have seen that an analysis of predefined feature subsets can provide new insights into brain feature discrimination. This expands the utility of a classification algorithm from a final output to an investigative process. The consideration of other data structures is equally conceivable: As an example, time series usually offer the possibility to divide them into subsets. On the one hand, there are multivariate time series, which represent several signals over time. Medical examples are the electrocardiogram (ECG) or

electroencephalography (EEG), where information is obtained from various leads or multiple electrodes on the scalp. If we build a machine learning model that extracts features from each signal independently (e.g., sliding-window approaches), we are able to describe the information gain of every signal. With this knowledge, the cardiologist's focus could be directed to a specific lead in the ECG - or the neurologist to a specific EEG-channel. On the other hand, a subdivision into phases is also conceivable for univariate time series.

However, it is precisely at this point that a limitation of our method becomes apparent. The features must be extracted independently of the subsequences. Modern end-to-end algorithms, such as convolutional neural networks, cannot be measured with the presented approach. Furthermore, a more in-depth analysis of the method based on a comparison to alternative algorithm (e.g., grouped SHAP values), different metrics, or multiclass problems are still pending for the future.

Nevertheless, the additional benefit of analyzing the importance of entire subsets through feature-based methods should not be ignored. In contrast to time series, all groupable types of data are conceivable, whether a questionnaire according to categories or a clinical examination according to organ parameters.

5. Conclusion

We presented the grouped permutation importance to achieve a better understanding of the underlying problem by examining subsets of features. In this context, better localization of information in the data produces new insights. Through an example from brain research, we have shown how different brain regions are involved in a decision-making process depending on the target variable. In addition, many applications are conceivable since, especially in medicine, data sets usually consist of a large number of characteristics that can be grouped together in clinically meaningful categories. Determining their impact on a predictive algorithm opens entirely new possibilities in understanding, detecting, and treating diseases.

References

[1] Varghese J. Artificial Intelligence in Medicine: Chances and Challenges for Wide Clinical Adoption. Visceral medicine, 36, 443-449, 2020.
[2] Breiman L. Random Forests. Machine Learning, 45(1), 5-32, 2001.
[3] Gregorutti B, Michel B, Saint-Pierre P. Grouped variable importance with random forests and application to multiple functional data analysis. Computational Statistics and Data Analysis, 90, 15-35, 2015.
[4] Loughrey J, Cunningham P. Overfitting in Wrapper-Based Feature Subset Selection: The Harder You Try the Worse it Gets. Research and Development in Intelligent Systems XXI, 33-43, 2005.
[5] Marcus DS, Wang TH, Parker J, Csernansky JG, Morris JC, Buckner RL. Open Access Series of Imaging Studies (OASIS): Cross-sectional MRI Data in Young, Middle Aged, Nondemented, and Demented Older Adults. Journal of Cognitive Neuroscience, 19, 1498-1507, 2007.
[6] Abraham A, Pedregosa F, Eickenberg M, Gervais P, Mueller A, Kossaifi J, et al. Machine learning for neuroimaging with scikit-learn. Frontiers in Neuroinformatics, 8:14, 2014.
[7] Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research, 12, 2825-2830, 2011.

Challenges of Trustable AI and Added-Value on Health
B. Séroussi et al. (Eds.)

doi:10.3233/SHTI220407

Machine Learning in Medicine: To Explain, or Not to Explain, That Is the Question

Cedric BOUSQUET[a,1] and Diva Beltramin[b]
[a]*Sorbonne Université, Inserm, université Paris 13, Laboratoire d'informatique médicale et d'ingénierie des connaissances en e-santé, LIMICS, F-75006 Paris, France*
[b]*Service de santé publique et information médicale, CHU de Saint Etienne, France*

Abstract. In 2022, the Medical Informatics Europe conference created a special topic called "Challenges of trustable AI and added-value on health" which was centered around the theme of eXplainable Artificial Intelligence. Unfortunately, two opposite views remain for biomedical applications of machine learning: accepting to use reliable but opaque models, vs. enforce models to be explainable. In this contribution we discuss these two opposite approaches and illustrate with examples the differences between them.

Keywords. Artificial intelligence, Explainability, Algorithms, Decision making, Trust

1. Introduction

Artificial intelligence systems are not yet used significantly in practice, which raises the question of their perceived usefulness [1], trustworthiness [2] and acceptance [3] by physicians. In 2022, the Medical Informatics Europe conference created a special topic called "Challenges of trustable AI and added-value on health" which was centered around the theme of eXplainable Artificial Intelligence [4]. The main idea is that new AI systems operate like black boxes [5] and cannot explain their predictions. We agree that this raises problems for the acceptance of such systems by physicians. However, imposing explainability as a mandatory criterion could lead to not being able to use the most efficient algorithms in certain situations. After presenting arguments in favor and then against explainibility of black boxes, we show that opinions are less clear-cut than it seems at first glance.

2. Results and discussion

Some models present inherent explainability such as decision trees where the explanation is the model. In applications on medical imaging, genomic data or hospital reports, data and models are very complex. The relationship between inputs and outputs is not linear and cannot be explained by a simple model, which requires post-hoc interpretation [6]. One solution is to build, in parallel with the black box, a model for which we know the

[1] Corresponding Author, Dr Cedric Bousquet, SSPIM, Bâtiment CIM42, chemin de la Marandière, Hôpital Nord, 42055 Saint Etienne; E-mail: cedric.bousquet@chu-st-etienne.fr.

execution logic, for example a linear regression model, by trying to make its response correspond as well as possible to the response of the black box. However, the substitution model presents lower performances than those of the black box. In the case of medical imaging, this is generally performed using saliency maps [7]. Post-hoc interpretations of black box predictions generate explanations that are not necessarily related to these predictions. This can lead to an erroneous belief that we understand the predictions of the model when we do not know its internal behavior [8].

Additionally, these explanation algorithms lack robustness because they may not generate similar explainations when confronted to similar data, as a small change in data may lead to different explanations. It is also difficult to retrace the steps when the system makes an error using post-hoc interpretation because it imperfectly approximates the function leading to the prediction. Saliency maps applied to images make it possible for the clinician to see where the regions of interest are, but do not indicate what is the characteristic that makes the regions interesting for the system [6]. Moreover, Abedayo et al. showed that the ability to evaluate predictions through a visual representation can mislead the user [9]. Experimentally, they established that some salience methods are independent of the model and the way the data is generated. Explainability is primarily intended for the system's implementers [6] who can thus determine why their system is not providing the correct answer, and thus improve it.

In some cases, it is truly impossible to produce an explanation that makes sense to the user. An example of this is DeepGestalt, an algorithm capable of recognizing facial phenotypes associated with genetic disorders; the system works well, however it is unable to give explanations to the end user of what were the features that induced the final diagnosis [10]. Explainability is certainly desirable and, although we hope that it will be implemented in future algorithms by design, it should not be constrained to the detriment of the use of AI systems that are difficult to explain but which can be of service to the patient.

References

[1] Hah H, Goldin DS. How Clinicians Perceive Artificial Intelligence-Assisted Technologies in Diagnostic Decision Making: Mixed Methods Approach. J Med Internet Res. 2021;23(12):e33540. doi: 10.2196/33540.

[2] Markus AF, Kors JA, Rijnbeek PR. The role of explainability in creating trustworthy artificial intelligence for health care: A comprehensive survey of the terminology, design choices, and evaluation strategies. J Biomed Inform. 2021;113:103655. doi: 10.1016/j.jbi.2020.103655.

[3] Bates DW, Kuperman GJ, Wang S et al. Ten commandments for effective clinical decision support: making the practice of evidence-based medicine a reality. J Am Med Inform Assoc. 2003;10(6):523-30.

[4] Belle V, Papantonis I. Principles and Practice of Explainable Machine Learning. Front Big Data. 2021;4:688969. doi: 10.3389/fdata.2021.688969. eCollection 2021.

[5] Kundu S. AI in medicine must be explainable. Nat Med. 2021;27(8):1328.

[6] Ghassemi M, Oakden-Rayner L, Beam AL. The false hope of current approaches to explainable artificial intelligence in health care. Lancet Digit Health. 2021;3(11):e745-e750. doi: 10.1016/S2589-7500(21)00208-9.

[7] Salahuddin Z, Woodruff HC, Chatterjee A, Lambin P. Transparency of deep neural networks for medical image analysis: A review of interpretability methods. Comput Biol Med. 2021;140:105111.

[8] Babic B, Gerke S, Evgeniou T, Cohen IG. Beware explanations from AI in health care. Science. 2021;373(6552):284-286.

[9] Adebayo J, Gilmer J, Muelly M, Goodfellow I, Hardt M, Kim B. Sanity checks for saliency maps. arXiv preprint 2018. arXiv:1810.03292.

[10] Gurovich Y, Hanani Y, Bar O et al. Identifying facial phenotypes of genetic disorders using deep learning. Nat Med. 2019;25(1):60-64.

Challenges of Trustable AI and Added-Value on Health
B. Séroussi et al. (Eds.)

doi:10.3233/SHTI220408

Phenotyping of Heart Failure with Preserved Ejection Faction Using Health Electronic Records and Echocardiography

Morgane PIERRE-JEAN[a,b,1], Erwan DONAL[a,b], Marc CUGGIA[a,b] and Guillaume BOUZILLÉ[a,b]
[a] *CHU de Rennes, Cardiologie, France*
[b] *LTSI, Inserm, 1099, Rennes, France*

Abstract. Patients suffering from heart failure (HF) symptoms and a normal left ventricular ejection fraction (LVEF 50%) present very different clinical phenotypes that could influence their survival. This study aims to identify phenotypes of this type of HF by using the medical information database from Rennes University Hospital Center. We present a preliminary work, where we explore the use of clinical variables from health electronic records (HER) in addition to echocardiography to identify several phenotypes of patients suffering from heart failure with preserved ejection fraction. The proposed methodology identifies 4 clusters with various characteristics (both clinical and echocardiographic) that are linked to survival (death, surgery, hospitalization). In the future, this work could be deployed as a tool for the physician to assess risks and contribute to support better care for patients.

Keywords. machine learning, phenotyping, Heart failure, Health electronic records, echocardiography

1. Introduction

The number of patients who suffer from heart failure (HF) symptoms and a normal left ventricular ejection faction (LVEF $\geq$50%) is growing up [1,2]. This pathology is referred to HF with preserved ejection faction (HFpEF). However, it represents an heterogeneous syndrome with very different clinical phenotypes [1,3]. This study aims to use both the clinical variables available in health electronic records and echocardiographic parameters to classify patients suffering from HFpEF into groups who likely share similar physiological profiles. Here, we proposed an algorithm to classify the patients into phenogroups. Finally, we attempted to links these phenogroups to their follow-up. We study three survival outputs: death, surgery and admission in cardiology. Heading

[1] Corresponding Author, Morgane Pierre-Jean; E-mail: morgane.pierre-jean@chu-rennes.fr.

2. Methods

The number of patients who suffer from heart failure (HF) symptoms and a normal left ventricular ejection faction (LVEF $\geq$50%) is growing up [1,2]. This pathology is referred to HF with preserved ejection faction (HFpEF). However, it represents an heterogeneous syndrome with very different clinical phenotypes [1,3]. This study aims to use both the clinical variables available in health electronic records and echocardiographic parameters to classify patients suffering from HFpEF into groups who likely share similar physiological profiles. Here, we proposed an algorithm to classify the patients into phenogroups. Finally, we attempted to links these phenogroups to their follow-up. We study three survival outputs: death, surgery and admission in cardiology.

We performed a two-step cluster analysis to identify common characteristics among patients. As a first step, we performed Principal Component Analysis (PCA) following by a spectral clustering to the 10 first coordinates of the PCA. The follow-up information about the vital status of the patients was obtain from both data available from the data warehouse and from the National Institute of Statistics and Economic Studies. We matched firstname, surname, date of birth and town of birth when it was available. We collected stays into cardiology department from the data warehouse and CIM-10 codes were used to extract the surgery information. We conducted a survival analysis on death, surgery, and the admission to the cardiology department.

The final aim is to classify new patients, then we used supervised machine learning algorithms to predict the phenogroup that are defined by the clustering algorithm. We optimized two algorithms: SVM and Random forest. We split the cohort into two sets (train and test), the algorithms are trained on the first set and we evaluate there performance on the second one. Performance was evaluated by computing AUC ROC and accuracy.

3. Results

The final sample eligible included around 2500 patients with echocardiography and suffering from HFpEF. The variables included in the two-step cluster analysis were both clinical (13 variables) and echographic (17 variables). Four clusters were identified from the clustering algorithm. The groups were relatively well balanced with respectively: 753, 744, 519 and 545 patients.

From these clusters, we extract significant distinct variables between clusters. After performing survival analysis, we observed that clusters have significative different survival curves in particular for death.

Optimization of machine learning models has been conducted on training sets (75%) in addition to bootstrap. Then, we evaluated the models on the test sets. The performances of these two models are satisfying with an AUC upper than 0.97 and an accuracy upper than 0.92. We retain the SVM model that provides better performance on the two criteria to classify patients.

4. Discussion

The clinical variable extraction from health records is not perfect because of automatic extraction, and some variables are of poor quality or missing for a non-negligible number of patients. Even if, the cardiologists help us to perform quality control of data, we can improve this part by using NLP techniques.

Imputation of missing values was done using KNN- algorithm. Our algorithm could be tested with the dataset imputed by the mean of the median of the variables for examples. Then, we could determine which imputation method gives the better results.

The final step could be to build a score to predict the phenotype of patients or even their survival (death, hospitalization, surgery).

5. Conclusion

We develop a POC of machine learning model to predict the phenotype of patients suffering from HFpEF from both clinical and echocardiography data from data warehouse of Rennes University Hospital Center. The phenotyping of HFpEF could improve the characterization of patients, the definition of the most appropriate treatments, and the care pathways.

References

[1] Fayol A, Wack M, Livrozet M, Carves JB, Domengé O, Vermersch E, Mirabel M, Karras A, Le Guen J, Blanchard A, Azizi M, Amar L, Bories MC, Mousseaux E, Carette C, Puymirat E, Hagège A, Jannot AS, Hulot JS. Aetiological classification and prognosis in patients with heart failure with preserved ejection fraction. ESC Heart Fail. 2022 Feb;9(1):519-530. doi: 10.1002/ehf2.13717. Epub 2021 Nov 29. PMID: 34841727; PMCID: PMC8788026.

[2] van Riet EE, Hoes AW, Wagenaar KP, Limburg A, Landman MA, Rutten FH. Epidemiology of heart failure: the prevalence of heart failure and ventricular dysfunction in older adults over time. A systematic review. Eur J Heart Fail. 2016 Mar;18(3):242-52. doi: 10.1002/ejhf.483. Epub 2016 Jan 4. PMID: 26727047.

[3] Galli E, Bourg C, Kosmala W, Oger E, Donal E. Phenomapping Heart Failure with Preserved Ejection Fraction Using Machine Learning Cluster Analysis: Prognostic and Therapeutic Implications. Heart Fail Clin. 2021 Jul;17(3):499-518. doi: 10.1016/j.hfc.2021.02.010. PMID: 34051979.

Challenges of Trustable AI and Added-Value on Health
B. Séroussi et al. (Eds.)

doi:10.3233/SHTI220409

Towards a Generic Description Schema for Clinical Decision Support Systems

Stefan VOGEL[a,1] and Dagmar KREFTING[a]
[a] *Department of Medical Informatics, University Medical Center Göttingen, Germany*

Abstract. Many clinical decision support systems (CDSS) have shown good performance in the research context, but only a few have been brought into routine care. Analysis of success or failure factors in the design of CDSS may support translation from development to routine care by guiding CDSS design and development along these factors. In this work, we propose a schema to describe CDSS designs in a consistent way. We focus on design criteria with the aim to investigate the observed translation gap in CDSS. Existing description models on different aspects relevant for CDSS are combined to a comprehensive schema that allows description and comparison of CDSS without limitation to the domain or architecture.

Keywords. Clinical Decision Support Systems, Artificial Intelligence, Harmonization, Software Architecture

1. Introduction

CDSS approaches have been under development for nearly seven decades [1]. Contrary to the manifold of CDSS developments, just a few of these systems have reached routine care. The CDSS software life cycle is complex and the systems are heterogeneous [2]. We propose a schema to describe CDSS at a detail level that is able to catch the relevant design factors without limiting its applicability. The overall aim of this work is to contribute to the research on success and failure factors of CDSS by consistent and comparable description of such systems.

2. Methods

To cover the main design factors of CDSS, existing description models and schemata for the topics “clinical decision-making”, “CDSS architecture” and “AI” are considered. The identified criteria are grouped according to the CDSS software life cycle components (a) initial design phase, (b) software, (c) inference and (d) knowledge base [2]. In the *clinical decision-making* process, the ProActIve concept is considered as a "business process" from which nine essential tasks can be identified that lead to a robust decision. [3]. For the topic CDSS architecture, different CDSS models [1, 4, 5] as well as an real CDSS implementation [6] were compared to identify core functional components and designs. For the topic AI the periodic system according to Hammond was considered [7].

[1] Corresponding author, Stefan VOGEL, University Medical Center Göttingen, Department of Medical Informatics, D-37075 Göttingen, Germany; E-mail: stefan.vogel@med.uni-goettingen.de.

3. Results

The research results gathered from the investigations of the three topics were summarized into the following schema as a generic description approach for CDSS.

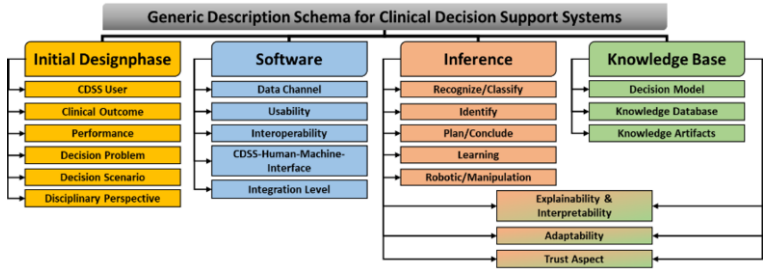


Figure 1. Generic description schema for CDSS

The schema allows to describe a CDSS design focusing on a specific CDSS life cycle component or to describe and evaluate the CDSS as a whole.

4. Discussion

The research in the field of CDSS and AI will continue to be dynamic. Therefore, it cannot be excluded that further components or criteria will have to be added to the schema in the future.

5. Conclusions

The generic description schema is a harmonization approach, which includes essential components and design criteria of CDSS.

References

[1] Wright A, Sittig DF. A four-phase model of the evolution of clinical decision support architectures. Int J Med Inform. 2008; 77(10):641-649. doi:10.1016/j.ijmedinf.2008.01.004

[2] Richter J, Vogel S. Illustration of Clinical Decision Support System Development Complexity. Stud Health Technol Inform. 2020 Jun 26; 272:261-264. doi: 10.3233/SHTI200544. PMID: 32604651.

[3] Hunink M, Weinstein MC, Wittenberg E, Drummond MF, Pliskin JS, Wong JB, Glasziou PP. Decision Making in Health and Medicine: Integrating Evidence and Values, 2nd Ed. Cambridge University Press, 2014, pp 5-26

[4] Zikos D, DeLellis N. CDSS-RM: a clinical decision support system reference model. BMC Med Res Methodol. 2018 Nov 16; 18(1):137. doi: 10.1186/s12874-018-0587-6.

[5] Boersch I, Heinsohn J, Socher R. Wissensverarbeitung – Eine Einführung in die Künstliche Intelligenz für Informatiker und Ingenieure 2nd Ed. Spektrum Akademischer Verlag ELSEVIER, 2007, pp 8-10

[6] Vogel S, Richter J, Wache S, Pischek-Koch K, Auchter S, Zebbities S, Güttler K, Hübner U, Pryzsucha M, Hüsers J, Sellemann B. Evaluation of a Clinical Decision Support System in the Domain of Chronic Wound Management. Stud Health Technol Inform. 2021 May 27;281:535-539

[7] Bauer W, Vocke C. (2020) Work in the Age of Artificial Intelligence – Challenges and Potentials for the Design of New Forms of Human-Machine Interaction. In: Kantola J, Nazir S (eds). Advances in Human Factors, Business Management and Leadership. AHFE 2019. Advances in Intelligent Systems and Computing, vol 961. Springer, Cham. https://doi.org/10.1007/978-3-030-20154-8_45

Challenges of Trustable AI and Added-Value on Health
B. Séroussi et al. (Eds.)

doi:10.3233/SHTI220410

Applying Artificial Intelligence Privacy Technology in the Healthcare Domain

Abigail GOLDSTEEN[a,1], Ariel FARKASH[a], Micha MOFFIE[a] and Ron SHMELKIN[a]
[a] *IBM Research – Haifa, Haifa, Israel*

Abstract. Regulations set out strict restrictions on processing personal data. ML models must also adhere to these restrictions, as it may be possible to infer personal information from trained models. In this paper, we demonstrate the use of two novel AI Privacy tools in a real-world healthcare application.

Keywords. GDPR, Privacy, Machine Learning, Artificial Intelligence, Healthcare

1. Introduction

There is a known tension between the need to analyze personal data, and the need to preserve privacy of data subjects, especially in the health domain. Data protection regulations, such as GDPR, set out strict restrictions on the collection and processing of personal data. As personal information may be derived from machine learning (ML) models using inference attacks [1], models must also adhere to these requirements[2].

Many techniques have been developed recently for privacy in ML models. However, few of them have been applied in real-world settings. We demonstrate the use of two such tools in a real system for early detection of melanoma[3]: ML model anonymization and data minimization. We demonstrate their incorporation into the system data flow and architecture and show initial results on a representative medical dataset.

2. Methods

First, model-guided anonymization [2] of the training data is performed and the model is retrained on the anonymized data. Data minimization [3] is then applied to reduce the amount and granularity of newly collected data input to the model. Both tools are available in the open-source AI Privacy Toolkit[4].

The melanoma diagnostic system is presented in Figure 1. Model anonymization is applied as part of the model training phase. It does not require changes to the training procedure itself, just an extra step of anonymizing the training data and retraining the model on the anonymized data. Data minimization is applied on the final model, during or after model validation. The resulting generalizations are fed into the data management component so it can filter and generalize features before uploading to the cloud.

[1] Corresponding Author, Abigail GOLDSTEEN, IBM Research – Haifa; E-mail: abigailt@il.ibm.com.
[2] https://www.europarl.europa.eu/thinktank/en/document/EPRS_STU(2020)641530
[3] https://itobos.eu/index.php
[4] https://github.com/IBM/ai-privacy-toolkit

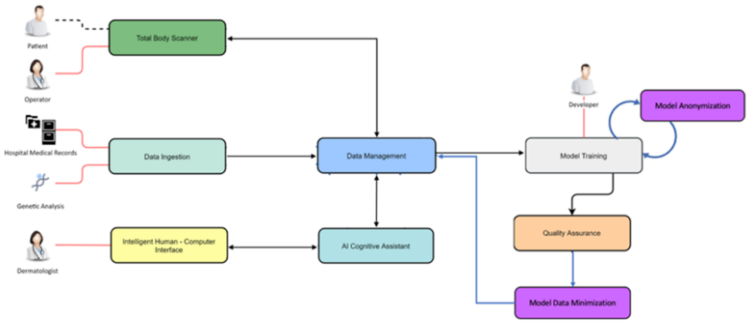


Figure 1. Melanoma diagnostic tool high-level architecture

3. Initial Results

The project has not yet begun collecting data, so results are shown on a similar dataset[5].

Table 1. Result of applying model-guided anonymization on the Hepatitis dataset with different values of k. Attack accuracy denotes the accuracy of a black-box membership inference attack on the resulting model.

Model type	k	Model accuracy	Attack accuracy
Random forest	None (original data)	0.81	0.66
	5	0.81	0.45
	10	0.78	0.46
Naïve Bayes	None (original data)	0.64	0.59
	5	0.78	0.46
	10	0.77	0.46

Table 2. Result of applying data minimization on the same dataset and models. Relative model accuracy denotes the accuracy of the model on the generalized data relative to its accuracy on the original data.

Model type	Relative model accuracy	Suppressed features
Random forest	0.98	Albumin
Naïve Bayes	0.93	Antivirals, Histology

Acknowledgements

This work has been supported by the iToBoS project funded by the European Union's Horizon 2020 research and innovation programme, under grant agreement No 965221.

References

[1] Fredrikson M, Lantz E, Jha S, Lin S, Page D, Ristenpart T. Privacy in Pharmacogenetics: An {End-to-End} Case Study of Personalized Warfarin Dosing. In23rd USENIX Security Symposium (USENIX Security 14) 2014 (pp. 17-32).

[2] Goldsteen, A., Ezov, G., Shmelkin, R., Moffie, M., Farkash, A. Anonymizing Machine Learning Models, In: Proceedings of the International Workshop on Data Privacy Management, 2021.

[3] Goldsteen A, Ezov G, Shmelkin R, Moffie M, Farkash A. Data minimization for GDPR compliance in machine learning models. AI and Ethics. 2021 Sep 26:1-5..

[5] https://archive.ics.uci.edu/ml/datasets/hepatitis

Challenges of Trustable AI and Added-Value on Health
B. Séroussi et al. (Eds.)

doi:10.3233/SHTI220411

Explainable Artificial Intelligence in Ambulatory Digital Dementia Screenings

Markus SCHINLE[a,1], Christina ERLER[a], Maximilian HESS[a] and Wilhelm STORK[b]
[a] *FZI Research Center for Information Technologies, Germany*
[b] *KIT Karlsruhe Institute of Technology, Germany*

Abstract. Recently, digital apps have entered the market to enable the early diagnosis of dementia by offering digital dementia screenings. Some of these apps use Machine Learning (ML) to predict cognitive impairment. The aim of this work is to find explanations for the predictions of such a mobile application called DemPredict using methods from the field of Explainable Artificial Intelligence (XAI). In order to evaluate which method is best suited, different XAI approaches are used and compared. However, the comparability of the results is a key challenge. By evaluating the trustworthiness, stability, and computation time of the methods, it is possible to identify the optimal XAI approaches for the respective algorithms.

Keywords. explainable artificial intelligence, interpretable machine learning, dementia screening, precision care, Alzheimer's disease

1. Introduction and Methods

Research shows that diagnosing dementia as early as possible is critical to the success of interventions [1]. This is also the aim of a digital health application called DemPredict using various psychometric tests as well as automated and optimized outcome of the assessment with ML [2,3]. To be successfully deployed in the outpatient setting, confidence in the results must be improved through explanations. To this end, for the best performing algorithms Random Forest (RF) and k-Nearest Neighbors (kNN), the extent to which XAI can be used will be examined [3]. From a structured literature review, it was decided to evaluate the model-agnostic procedures LIME and SHAP [4] as well as the model-specific procedures CHIRPS [5] for RF and an example-based explanation of the kNN. Since there is no common consensus on what explainability means for ML, there are no uniform performance measures against which an explanation can be evaluated [4]. Doshi-Velez and Kim formulated a general approach to evaluate the different XAI methods by distinguishing between different levels [6]. In this work, the functional level evaluation is focused. It is based on the three criteria: computation time, trustworthiness, and stability. The first criterion is intended to measure the runtime required by a XAI method to generate an explanation. To obtain a robust value, this measurement is performed for ten test samples, with the average of the calculation times serving as a comparison value. The trustworthiness of a XAI method is evaluated by measuring how true the explanations of this procedure are [7]. In order to quantify the trustworthiness of the XAI method, the Most Relevant First (MoRF) perturbation curve

[1] Corresponding Author, Markus Schinle; E-mail: schinle@fzi.de.

is calculated [8]. This measures how the classification probabilities change if the information is removed from the input of the ML algorithm [8] and can be implemented by removing features and replacing the values of selected features with the mean value of the corresponding class. The stability property guarantees that the same explanations are generated under the same conditions [9]. In order to measure the stability of a method, it is examined how the XAI methods behave when generating explanations of the same sample several times [9,10]. The coefficient of variation is calculated to measure the dispersion of the influences of a feature.

2. Results, Discussion and Conclusions

In terms of computation times, LIME shows the fastest computation times on both ML models. CHIRPS lies between LIME and SHAP, while the example-based explanation of the kNN has the highest time on average. Regarding the trustworthiness of the model-agnostic procedures, SHAP shows lower MoRF values than LIME on both ML models. Lower MoRF values indicate higher trustworthiness in the explanations, as the classification probability decreases faster if the important features are replaced with the mean. In comparison to the model-agnostic procedures, CHIRPS does not calculate the influence values of the features but generates a classification rule as an explanation. Instead of the MoRF curve, the evaluation of trustworthiness is measured using the performance measures of the classification rule. These show comparable performance to those of the underlying RF, meaning that the most important components of the RF have been identified. For the example-based explanations, trustworthiness can be considered optimal, since the explanations correspond to the k nearest neighbors, which are also used for the classification. As CHIRPS and the example-based explanations are based on deterministic algorithms, the stability of the methods can be considered optimal. Among the model-agnostic methods, in particular, LIME shows instabilities. In contrast, the SHAP method shows only minimal fluctuations in the coefficient of variation. A concept was developed to carry out an evaluation that offers the possibility to compare conceptually different XAI methods on the functional level. However, the model-agnostic methods can be compared in a more differentiated way, as the MoRF curve and the stability can be calculated explicitly.

References

[1] Livingston G, et al. Dementia prevention, intervention, and care. Lancet. 2017;390(10113):2673-4.
[2] Schinle M, et al. An Approach to digitalize Psychological Tests to support Diagnosis of Alzheimer's Disease in Ambulatory Care. In 2018 IEEE MeMeA 2018 Jun 11 (pp. 1-6). IEEE.
[3] Schinle M, et al. Data-driven Development of Digital Health Applications on the Example of Dementia Screening. In 2021 IEEE MeMeA 2021 Jun 23 (pp. 1-6). IEEE.
[4] Molnar. Interpretable machine learning. A Guide for Making Black Box Models Explainable. 2019.
[5] Hatwell J. CHIRPS: Explaining random forest classification. Artific Intell Review. 2020;53:5747-88.
[6] Doshi-Velez F, Kim B. Towards a rigorous science of interpretable machine learning. 2017.
[7] Alvarez D, Jaakkola T. Towards robust interpretability with self-explaining neural networks. 2018.
[8] Samek W, et al. Evaluating the visualization of what a deep neural network has learned. IEEE transactions on neural networks and learning systems. 2016 Aug 25;28(11):2660-73.
[9] Visani G, et al. Statistical stability indices for LIME: Obtaining reliable explanations for machine learning models. Journal of the Operational Research Society. 2022 Jan 2;73(1):91-101.
[10] Shankaranarayana SM, Runje D. ALIME: Autoencoder based approach for local interpretability. 2019.

Challenges of Trustable AI and Added-Value on Health
B. Séroussi et al. (Eds.)

doi:10.3233/SHTI220412

Behavioral Segmentation for Enhanced Peer-to-Peer Patient Education

Alessia PAGLIALONGA[a,1], Rebecca THEAL[b], David BARBER[b], Robert KYBA[c], Aziz GUERGACHI[d, e], and Karim KESHAVJEE[f]
[a] *National Research Council of Italy (CNR) IEIIT Institute, Milan, Italy*
[b] *Department of Family Medicine, Queen's University, Kingston, Canada*
[c] *Strategic Global Counsel, Toronto, Canada*
[d] *Ryerson University, Ted Rogers School of Management, Toronto, Canada*
[e] *York University, Department of Mathematics and Statistics, Toronto, Canada*
[f] *Institute of Health Policy, Management and Evaluation, University of Toronto, Canada*

Abstract. The aim of this study was to develop a peer-to-peer virtual intervention for patients with type 2 diabetes from different segments: patients who take several medications (medication group), patients who do not take diabetes medications (lifestyle group), and a mixed group. Preliminary results showed that patients in the lifestyle group were interested in preventive strategies, reporting better learning experience and higher motivation than those in the medication group. Future research is needed to design approaches tailored to patients in the medication group.

Keywords. behavioral segmentation, peer-to-peer support, type 2 diabetes

1. Background

Type 2 Diabetes (T2D) is a major challenge for health systems worldwide [1]. A minority of patients achieve good control of disease whereas the majority struggle to do so, despite having access to the same resources [2]-[4]. Not all patients have the same preferences when it comes to treatments of T2D. The aim of this study is to develop a new, scalable, data driven approach for patients with T2D in the form of peer-to-peer support in a virtual environment, based on patient segmentation on medication-taking behavior. Segmentation is widely used in marketing [5] but still largely unexplored in healthcare [6],[7] and it is usually not adopted for the provision of tailored interventions, despite growing evidence that it could support patient self-management [8],[9].

2. Pilot Experiment

Three peer-to-peer virtual workshops were performed (N=14 patients out of a dataset of 825 patients): medication group (patients taking anti-hypertensives, anti-cholesterol, 1st line, and 2nd line diabetes drugs); lifestyle group (patients not taking diabetes drugs), and

[1] Corresponding Author, Alessia Paglialonga, Institute of Electronics, Information Engineering and Telecommunications (IEIIT), Consiglio Nazionale delle Ricerche (CNR), piazza Leonardo da Vinci 32, 20133 Milan, Italy; E-mail: alessia.paglialonga@ieiit.cnr.it.

a mixed group. Patients filled out surveys regarding attitudes, learning and motivation. The protocol was approved by the Ethics Boards at Queen's University Health Sciences & Affiliated Teaching Hospitals (Oct 5, 2020) and at Ryerson University (Mar 10, 2021).

The groups did not differ in terms of age and gender distribution, and the average values of biomarkers (glycated hemoglobin, low-density lipoprotein, systolic and diastolic blood pressure) were similar. All participants appreciated the value of being with other people with the same disease to share their stories, regardless of the patient segment and regardless of their level of control of disease. Differences were observed between groups in terms of self-reported learning and preferred topics (i.e., diet, exercise, and management of diabetes related symptoms in the mixed and lifestyle groups). In the mixed and lifestyle groups, patients identified specific practical solutions, they were satisfied with the format, and they were willing to recommend the workshop to others, whereas patients in the medication group were interested in managing the 'effects' of diabetes (symptoms, stress, mental health) rather than in preventive strategies.

Future research will be needed to develop approaches tailored to the needs of patients in the medication group (e.g., a discussion oriented to managing the effects of diabetes and improving medication adherence). The proposed approach is potentially scalable as it uses a fully virtual process, it does not require healthcare provider supervision, and is based on EMR data that are widely used globally.

Acknowledgements

This study was partially supported by Nesta through project SCI-PHI, Segment-based Collective Intelligence for Population Health Improvement. Nesta's collective intelligence grants programme is supporting experiments that combine human and machine intelligence to solve social problems. The £500,000 fund is co-funded by Nesta, Wellcome Trust, Cloudera Foundation and Omidyar Network.

References

[1] Aziz Z, Riddell MA, Absetz P, Brand M, Oldenburg B. Peer support to improve diabetes care: an implementation evaluation of the Australasian Peers for Progress Diabetes Program. BMC Public Health. 2018;18(1).

[2] Alyami M, Serlachius A, O'Donovan CE, Werf B, Broadbent E. A systematic review of illness perception interventions in type 2 diabetes: Effects on glycaemic control and illness perceptions. Diabet Med. 2020;38(3).

[3] Sanal TS, Nair, NS, Adhikari P. Factors associated with poor control of type 2 diabetes mellitus: A systematic review and Meta-analysis. Journal of Diabetology 2011;3(1):1-10.

[4] Theal R, Paglialonga A, Kyba R, Barber D, Guergachi A, Keshavjee K. Applying Patient Segmentation in Primary Care to Develop Peer-to-Peer Support Groups for Patients with Type 2 Diabetes. 2021 Trillium Primary Health Care Research Day, Oct 15, 2021, virtual conference.

[5] Tuckwell KJ, Jaffey M (2014). Market Segmentation and Target Marketing. In: THINK Marketing, 2nd Edition, 2014. Chapter 6.

[6] Chong JL, Lim KK, Matchar DB. Population segmentation based on healthcare needs: a systematic review. Syst Rev. 2019;8(1).

[7] Patient Segmentation: Adjust the Production Logic to the Medical Knowledge Applied and the Patient's Ability to Self-Manage-A Discussion Paper. Front Public Health. 2020;8.

[8] Paglialonga A, Patel AA, Pinto E, Mugambi D, Keshavjee K. The Healthcare System Perspective in mHealth. m_Health Current and Future Applications. Springer, Cham 2019:127-142.

[9] Vuik SI, Mayer EK, Darzi A. Patient Segmentation Analysis Offers Significant Benefits for Integrated Care and Support. Health Affairs. 2016;35(5):769-775.

Challenges of Trustable AI and Added-Value on Health
B. Séroussi et al. (Eds.)

doi:10.3233/SHTI220413

A Confidence Interval-Based Method for Classifier Re-Calibration

Andrea CAMPAGNER [a,1] and Lorenzo FAMIGLINI [a] and Federico CABITZA [a,b]
[a] *University of Milano-Bicocca, Milano, Italy*
[b] *IRCCS Istituto Ortopedico Galeazzi, Milan, Italy*

Abstract. We propose a re-calibration method for Machine Learning models, based on computing confidence intervals for the predicted confidence scores. We show the effectiveness of the proposed method on a COVID-19 diagnosis benchmark.

Keywords. Calibration, confidence interval, medical ML, trustable AI

1. Introduction

Machine Learning (ML) has shown promising accuracy in the clinical domain, but there are important limitations in terms of other quality dimensions [1], including *calibration* [4]: this is the extent the confidence scores associated with each prediction are close to the observed frequency of events. A possible solution to this problem is to apply a re-calibration method, so to adjust confidence scores [3]. Several such techniques have been proposed in the literature; however, existing methods do not provide any guarantee and require additional data for re-calibration.

2. Method

In this article we propose a re-calibration method based on the computation of confidence intervals for the confidence scores provided by any ML model. Let $S = \{(x_i, y_i)\}_{i=1}^{n}$ be the training set, and h a classifier, where $h(x)$ is the confidence score associated the positive class. We first partition S into k bins $S_k^1, \ldots, S_k^k$, by sorting the instances' confidence scores. Then, this partition is used to compute confidence intervals around $h(x)$: given a confidence score $h(x)$ falling in a bin S_k^i, we compute a confidence interval (given $\alpha \in (0,1)$) as: $\left[\max\{0, h(x) - \frac{\sqrt{2\hat{\sigma}_i} \cdot erf^{-1}(\alpha)}{|S_k^i|+1}\}, \min\{1, h(x) + \frac{\sqrt{2\hat{\sigma}_i} \cdot erf^{-1}(\alpha)}{|S_k^i|+1}\}\right]$, where $h(x) \in S_k^i$, $\hat{\sigma}_i$ is the average confidence score in bin S_k^i, and erf is the error function.

We evaluated the proposed method, in comparison with other re-calibration methods, on a public dataset for the task of COVID-19 diagnosis from routine blood exams [1]. The training set consists of 1736 samples collected from February to May 2020 at the IRCCS Ospedale San Raffaele (OSR), in Milan, Italy. The test set consists of 224 sam-

[1]Corresponding Author: Andrea Campagner, a.campagner@campus.unimib.it

ples collected in November 2020 at IRCCS OSR. See [1] for further details. In regard to ML models, we used a state-of-the-art SVM-based model [1] as baseline and compared 4 re-calibration methods: Sigmoid regression (SR); Isotonic regression [3] (IR); Venn prediction [2] (VP); our proposed method ($\alpha = 0.90$). Models were compared in terms of the Brier score $\frac{1}{n}\sum_{i=1}^{n}(h(x_i) - y_i)^2$ and graphical analysis of the reliability diagrams.

3. Results and Discussion

The results of the experiment are shown in Figure 1. The proposed method reported a consistent improvement in calibration, as shown by the lower Brier score and the fact that the bisector line lies within the interval bounds. By contrast, the other methods did not provide any improvement compared to the baseline model, in terms of Brier score.

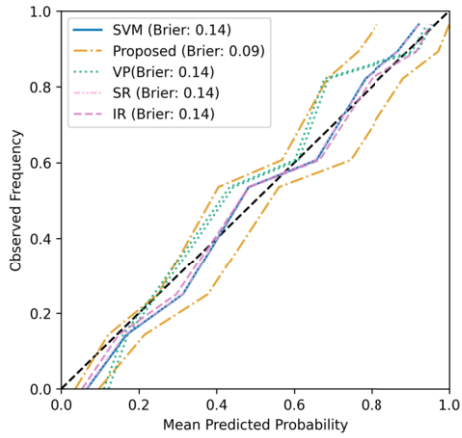


Figure 1. Reliability diagram for the considered models. The dashed line denotes perfect calibration.

4. Conclusion

We proposed a novel re-calibration method, based on computing confidence intervals for the confidence scores provided by a ML model. Through an illustrative experiment, we showed that the proposed technique provides better calibration than existing methods. However, further and more extensive experimental validation should be conducted.

References

[1] F. Cabitza, A. Campagner, et al. The importance of being external. Methodological insights for the external validation of machine learning models in medicine. *Comput Methods Programs Biomed*, 208:106288, 2021.

[2] A. Lambrou, I. Nouretdinov, et al. Inductive venn prediction. *Ann Math Artif Intell*, 74(1):181–201, 2015.

[3] A. Niculescu-Mizil and R. Caruana. Predicting good probabilities with supervised learning. In *Proceedings of the 22nd ICML*, pages 625–632, 2005.

[4] B. Van Calster, D. J. McLernon, et al. Calibration: the achilles heel of predictive analytics. *BMC medicine*, 17(1):1–7, 2019.

Challenges of Trustable AI and Added-Value on Health
B. Séroussi et al. (Eds.)

doi:10.3233/SHTI220414

Physiotherapists' Views on the Software Monitoring Application of a Wearable Assistive Glove

Valentina BARTALI[a,1] and Vera BULSINK [a]
[a] *Roessingh Research & Development, Roessinghsbleekweg 33b, 7522 AH Enschede, The Netherlands*

Abstract. The FutureGlove project aims to further develop the CarbonHand system (a wearable assistive glove) by introducing a new software monitoring application. Within this study, we consulted five physiotherapists during online sessions to gain their views on use of data and patients' empowerment. These first insights suggested that more data needs to be collected than it is doing now to monitor the patients' progress, that therapists would like to personalize the amount of freedom they provide to their patients, and that they would like decision support for algorithms, but they do not fully trust them.

Keywords. Hand rehabilitation, patients' empowerment, trust in algorithms

1. Introduction

In FutureGlove, a project within the Eurostars Programme (project ID: 114821) the Carbonhand system is being further developed into a next generation light, wearable robotic glove for hand impaired patients. The system has a high potential to act as a therapeutic assistive device for support during active daily living (ADL) and during therapy. The Carbonhand showed improvement in grip, strength and improved hand function in clinical studies [1,2]. For more information on the CarbonHand system and the glove's settings, please refer to the website of BioServo: https://www.bioservo.com/healthcare.

A software monitoring application providing access to the patients' results is part of this new technology. These will enable the rehabilitation process while the patient has the freedom to use the device on his or her own. The development follows an iterative, user-centric approach, and ensures that the design is based on user's needs.

2. Method

Online design session interviews were conducted with five physiotherapists to acquire their insights on functionalities of the software monitoring application, specifically regarding use of data and patients' empowerment. The participants gave permission to

[1] Corresponding Author, Valentina Bartali, Roessingh Research & Development, Roessinghsbleekweg 36b, 7522 AH Enschede, the Netherlands; E-mail: v.bartali@rrd.nl.

be video and audio recorded, and to use their data. After transcribing the interviews, two cycles of inductive (thus whilst reading the interview) coding as described by Saldaña were performed, with a good intercoder reliability (Kappa coefficient of .82) [3].

3. Results and Discussion

According to the therapists, more data needs to be collected than the system is doing now [2], with the goal of monitoring the patients' progress. The most mentioned parameters are: glove' usage, number of grasp movements, and glove's strength the user needs to make a grip.

The power given to a patient depends on his or her cognition and technological level. Accordingly, therapists should be able to personalize the amount of freedom given to a patient. Furthermore, patients should always discuss the data with the therapist to avoid misinterpretation. However, not giving patients the full power to use the technology on their own contrasts with an emerging approach to eHealth where patient empowerment is used to promote self-management and patient independence [4]. A way for patients to use the glove more independently and empower them is to have an algorithm learning from the parameters retrieved and automatically change the glove's settings and give advice. Nevertheless, an algorithm cannot be fully trusted to do that, also because it cannot know patients' hand dysfunction. Knowing that could give more reliable suggestions, but it might still not be safe to use as there are many factors which cannot be predicted. This could harm the patient, as debated in [5].

However, as smart decision making by algorithms has high potential, could also save time for therapists, and create personalised messages for patients. Because of the benefits, it is important to make this advice trustworthy, therefor evidence-based, good quality and user-friendly presentation of data is needed [6]. Moreover, ethical considerations should be made [5].

Because of the low number of participants, general conclusions cannot be drawn. However, the results of this study lead to recommendations for the development of the first prototype of the software monitoring application of FutureGlove. Therapists and patients will be further involved during testing and the further refinement of the prototype.

References

[1] Radder B, Prange-Lasonder GB, Kottink A, Holmberg J, Sletta K, van Dijk M et al. Home rehabilitation supported by a wearable soft-robotic device for improving hand function in older adults: A pilot randomized controlled trial. PLoS ONE. 2019;14(8)
[2] Kottink AI, Nikamp CD, Buurke JH, Bos F, Sluis CK, den Broek MV, Prange-Lasonder GB. Six weeks Use of a Wearable Soft-robotic Glove During ADL: Preliminary Results of Ongoing Clinical Study. In International Conference on NeuroRehabilitation; 2020 Oct. Springer, Cham. pp. 15-20
[3] Saldaña J. The coding manual for qualitative researchers (2nd Ed.). Sage, 2012.
[4] Risling T, Martinez J, Young J, Thorp-Froslie N. Evaluating patient empowerment in association with eHealth technology: scoping review. Journal of Medical Internet Research, 2017;19(9), 329.
[5] Grote H, Berens, P. On the ethics of algorithmic decision-making in healthcare. J Med Ethics. 2020 Mar;46(3):205-211.
[6] Banzi R, Liberati A, Moschetti I, Tagliabue L, Moja L. A Review of Online Evidence-based Practice Point-of-Care Information Summary Providers. J Med Internet Res 2010;12(3):e26

Challenges of Trustable AI and Added-Value on Health
B. Séroussi et al. (Eds.)

doi:10.3233/SHTI220415

Meaningful Activity Replacement Recommendations in Dementia

Jonathan TURNER[a,1], Michael WILSON[b], Ciarán NUGENT[a], Damon BERRY[a], Julie DOYLE[b] and Dympna O'SULLIVAN[a]
[a] *Technological University Dublin*
[b] *Dundalk Institute of Technology*

Abstract. Exercise of meaningful activities is important for people living with dementia, both for quality of life and to maintain the necessary basic activities of daily living. A method is proposed for recommendation of replacements for lost meaningful activities that accounts for the need to maintain activities of daily living.

Keywords. Dementia; goal modelling; self-care; independent living

1. Introduction

Dementia is a set of symptoms that include deterioration in memory and a decline in the ability to perform everyday activities. Persons Living With Dementia (PLWDs) require increasing care as their underlying disease progresses. There are benefits of independent living to PLWDs in improved quality of life, and societal benefits of reduced costs compared to institutional care [1]. To maintain independent living, PLWDS need to continue to perform Activities of Daily Living (ADL) [2] and Instrumental Activities of Daily Living (IADL) [3]. Quality of life (QoL) is enhanced by Meaningful Activities (MA) which have the additional benefits of exercising (I)ADLs, thus slowing decline [4].

If the ability to perform MAs is lost, QoL is reduced; (I)ADLs and skills intrinsic to an MA may no longer be performed. For example, an MA of 'attending church' may exercise the 'dressing' and 'transportation' IADLs, and cognitive, language and social skills. Should loss of an MA occur, we seek to recommend alternative MAs that will maintain a PLWD's QoL and ensure that (I)ADLs and skills continue to be exercised

2. Methods

For each MA, we determine which other activities are exercised. Table 1 gives an example of some MAs and the activities and skills that each exercises. Should an MA no longer be performed, the set of MAs is used to generate a ranked recommendation list by using a simple k nearest neighbour (knn) method [5]. Detection and recommendation need not be 'real time' but a swift recommendation would be beneficial to help minimize loss of ability to perform the component activities of an MA.

[1] Corresponding Author, Department of Computer Science, Technological University Dublin, Dublin, Ireland; E-mail: jonathan.turner@tudublin.ie

Table 1. Example MAs and the ADLS and IADLs that they exercise.

	ADL	ADL	ADL	ADL	ADL	ADL	IADL	IADL	IADL	IADL	IADL	IADL	skill	skill	skill	skill	skill	skill
Activity_name	**Bathing**	**Dressing**	**Toileting**	**Transferring**	**Continence**	**Feeding**	**Telephone**	**Shopping**	**Food**	**Laundry**	**Transportation**	**Finances**	**Motor**	**Cognitive**	**Vision**	**Sensory**	**Language**	**Social**
Meals with family	N	Y	N	Y	N	Y	N	N	Y	N	N	N	Y	Y	Y	N	Y	Y
Coffee with friends	N	Y	N	Y	Y	Y	N	N	N	N	Y	N	Y	Y	N	Y	Y	Y
Current affairs	N	N	N	N	N	N	N	N	N	N	N	N	N	Y	N	N	Y	N
Sport (playing)	N	Y	N	Y	N	N	N	N	N	N	Y	N	Y	Y	Y	Y	N	N

3. Results

A simple example based on the information shown in Table 1 and using the simple knn method is given. If we assume that a PLWD can no longer play sport, we see how close the other MAs in our set are by simple matching of activities and skills. This gives us a ranked list of suggested replacement MAs: coffee with friends (6 matches), meals with family (5), current affairs (1). So, we would recommend consideration be given to 'coffee with friends' as an activity that could replace the benefits lost by no longer playing sport.

4. Discussion

We plan to further refine the replacement activity selection algorithm and the selection criteria, to include themes that describe the rewards of MAs such as physical activity, being part of a community, or spiritual rewards; and to include flags to filter out MAs that may be inappropriate to recommend, e.g. 'visiting graveyards'.

5. Conclusions

There is potential to help PLWDs maintain essential (I)ADLs and skills if performance of an MA that exercises (I)ADLs and skills declines, by generating a ranked list of suggested replacement MAs. By encouraging PLWDs to replace lost MAs by alternative MAs, it is suggested that the period of independent living for PLWDs may be extended.

References

[1] C. Olsen et al, "Differences in quality of life in home-dwelling persons and nursing home residents with dementia - a cross-sectional study," BMC Geriatr. 2016 Jul 11;16:137.

[2] S. Katz, A. B. Ford, R. W. Moskowitz, B. A. Jackson, and M. W. Jaffe, "Studies of illness in the aged. The Index of ADL: a standardized measure of biological and psychosocial function," JAMA 185, pp:914-919, 1963.

[3] M. P. Lawton and E. M. Brody, "Assessment of older people: Self-maintaining and instrumental activities of daily living," The Gerontologist, 1969 9(3), pp:179-186

[4] G. Rakesh, S.T. Szabo, G.S. Alexopoulos and A.S. Zannas, "Strategies for dementia prevention: latest evidence and implications," Ther Adv Chronic Dis. 2017 Aug;8(8-9):121-136.

[5] Fix, Evelyn; Hodges, Joseph L. (1951). Discriminatory Analysis. Nonparametric Discrimination: Consistency Properties (PDF) (Report). USAF School of Aviation Medicine, Randolph Field, Texas

Challenges of Trustable AI and Added-Value on Health
B. Séroussi et al. (Eds.)

doi:10.3233/SHTI220416

Instruments Segmentation in X-ray Fluoroscopic Images for Endoscopic Retrograde Cholangio Pancreatography

Garance MARTIN[a,1], Saad EL-MADAFRI[a], Aymeric BECQ[b,c,2], Jérôme SZEWCZYK[b] and Isabelle BLOCH[a]
[a] *Sorbonne Université, CNRS, LIP6, Paris, France*
[b] *Sorbonne Université, CNRS, INSERM, ISIR, Paris, France*
[c] *UPEC, Hôpital Henri Mondor, Service de Gastroentérologie, APHP, Créteil, France*

Abstract. In this work, we propose a method to segment endoscope and guidewire from 2D X-ray fluoroscopic images of an endoscopic retrograde cholangiopancreatography (ERCP). We used an improved U-Net model. We obtained a Dice score of 0.94±0.05 for endoscope segmentation and a Hausdorff distance of 24.26 pixels for the guidewire segmentation. These preliminary results pave the way for further applications aiming at aiding the medical procedure.

Keywords. ERCP, X-Ray Fluoroscopy, U-Net

1. Introduction

Endoscopic Retrograde CholangioPancreatography is a minimally invasive procedure allowing draining the bile ducts in the setting of stones or neoplastic structuring. The instruments and guidewire used are radio-opaque and can therefore be monitored within the bile ducts using 2D X-ray fluoroscopy. However, contrast agent should remain limited because of the risk of cholangitis. In light of these shortcomings, this preliminary work aims to segment instruments in 2D X-ray fluoroscopy images.

2. Methods

We used a well-known approach in biomedical imaging called U-Net [1] to segment the endoscope and the guidewire. To cope with the difficulties related to the topology of the guidewire which is long and thin, the low contrast and signal to noise ratio, we propose specific pre- and post-processing steps to enhance this deep learning method.
Endoscope segmentation. We followed the same steps of pre-processing as per the nnU-Net [2] approach, consisting in cropping to keep non-zero regions only, normalizing intensity and data augmentation. The image database contained 300 annotated images.

[1] Corresponding Author, Garance MARTIN; E-mail: garance.martin@lip6.fr
[2] Images were provided by the Gastroenterology and Endoscopy Division of Hôpital Saint-Antoine, Sorbonne Université, Paris, France

Data augmentation allowed for the creation of 3200 additional ones. We applied elastic transformations in a limited range to keep medically plausible images. A first dataset of 2100 images was used for training, a second of 700 images for validation, and a third of 700 images for test.
Training parameters. We trained a generic improved U-Net with a leaky ReLU, Batch normalization and a loss function combining Binary Cross-entropy and Dice loss.
Post-processing. We applied a post-processing to remove small artifacts from the network prediction. The post-processing consisted of a morphological opening with a structuring element of 10x10 pixels followed by a morphological closing by a structuring element of 70x70 pixels.
Guidewire segmentation. Some additional steps were added to the preprocessing. We enhanced the guidewire with a morphological erosion and dilated the binary mask (annotation) with a structuring element of 3x3 pixels. Then, we generated 1000 images from the 100 previously annotated images. Finally, we included a regularization term in the loss function to overcome large class imbalance. This term is based on the percentage of the guidewire in the images (less than 1%), as learned from the database, and constrains the model to prioritize guidewire pixels instead of background.

3. Results

The Dice score was used to evaluate the endoscope segmentation. After post- processing, an average Dice score of 0.94 0.05 was obtained. Furthermore, we applied segmentation activation map based on gradient [3] to visualize which pixels are used by the network during segmentation. The results clearly demonstrate that pixels from endoscope only are used for the prediction making. To evaluate the segmentation of the guidewire we used a stricter metric, the Hausdorff Distance (HD) which is more adapted to long, thin objects. The smaller the HD the better the segmentation. An average HD of 24.26 pixels for native images of size 1024x1024 pixels was obtained. Qualitative results are available at this link: https://github.com/garancemartin/mie_results.

4. Discussion & conclusion

The endoscope segmentation is very promising. More annotated data would certainly improve this result. On the other hand, although the proposed additional steps improve the performances on the guidewire segmentation, the results do not completely meet expectations for now.
In future work we plan to introduce hybrid AI, combining knowledge not only on the guidewire but also on the structure of the biliary ducts.

References

[1] Ronneberger O, Fischer P, Brox T. U-Net: Convolutional networks for biomedical image segmentation. In: MICCAI; 2015. p. 234-41.
[2] Isensee F, Jaeger PF, Kohl SAA, Petersen J, Maier-Hein KH. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. Nature Methods;18(2):203-11.
[3] Vinogradova K, Dibrov A, Myers G. Towards Interpretable Semantic Segmentation via Gradient-weighted Class Activation Mapping. In: AAAI; 2020.

Challenges of Trustable AI and Added-Value on Health
B. Séroussi et al. (Eds.)

doi:10.3233/SHTI220417

Implementing a Microservices Architecture for Predicting the Opinion of Twitter Users on COVID Vaccines

Guillaume GUERDOUX[a], Bissan AUDEH[c], Théophile TIFFET[b] and Cédric BOUSQUET[b,c,1]
[a] *Geegz, Paris, France*
[b] *Unit of Public health, University hospital of Saint-Etienne, France*
[c] *Sorbonne Université, INSERM, Univ Paris 13, Laboratoire d'Informatique Médicale et d'Ingénierie des Connaissances pour la e-Santé, LIMICS, 75006, Paris, France*

Abstract. A strong trend in the software industry is to merge the activities of deployment and operationalization through the DevOps approach, which in the case of artificial intelligence is called Machine Learning Operations (MLOps). We present here a microservices architecture containing the whole pipeline (frontend, backend, data predictions) hosted in Docker containers which exposes a model implemented for opinion prediction in Twitter on the COVID vaccines. This is the first description in the literature of implementing a microservice architecture using TorchServe, a library for serving Pytorch models.

Keywords. Artificial Intelligence, MLOps, COVID-19, Social Media, Vaccines

1. Introduction

The remarkable performance of deep learning and its ongoing improvements raises the question of its usability in real life in the medical context. In this paper, we evaluate the feasibility of implementing a microservices architecture for the deployment of a deep learning model to classify Twitter users' opinion about COVID-19 vaccination that was implemented in a previous work [1]. In a nutshell, a deep learning model was implemented with PyTorch and CamemBERT [2], a French variant of Bidirectional Encoder Representations from Transformers (BERT) [3].

2. Method

We implemented an architecture based on three components within Docker containers, NGINX and two microservices: a backend implemented with Django-uWSGI and a prediction application programming interface deployed with TorchServe. We customized an inference handler and started TorchServe to serve the model, listening for clients' requests, and processing these requests. Docker-compose was used to define,

[1] Corresponding Author, Dr Cedric Bousquet, SSPIM, Bâtiment CIM42, chemin de la Marandière, Hôpital Nord, 42055 Saint Etienne, France; E-mail: cedric.bousquet@chu-st-etienne.fr.

build and manage the containers providing the services. Communications between clients and the server are managed through a RESTful API without maintaining the session state, hence the use of a JSON web token each time the client requests a service requiring authentication. The architecture is shown in figure 1.

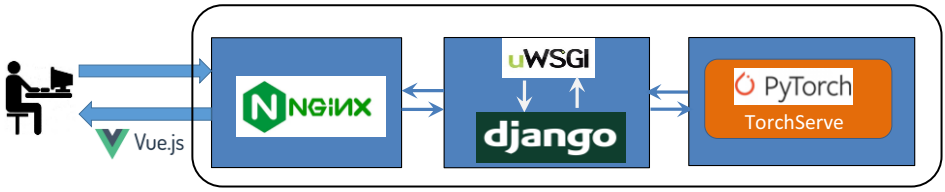


Figure 1. The microservices architecture.

3. Results and Discussion

As a proof of concept, we have implemented a simple web interface that allows the user to request the analysis of a sentence using our model. This request is then forwarded to Django via NGINX and uWSGI, then validated (format, authorization, etc.) and finally sent to our TorchServe service for prediction. Finally, the result is presented to the user in the same web interface. This interface was implemented using Vue.js, an open source JavaScript framework for building user interface and single-page applications. The F-Score of the classifier was 0.75 (precision: 0.74; recall: 0.75) [1].

This work demonstrates the feasibility of integrating a deep learning model with other applications once it is served using the proposed architecture. Furthermore, it is quite flexible, and it can be modified to meet various requirements. To our knowledge, this is the first description of a microservices architecture using TorchServe in the medical literature. This library presents two benefits: First, TorchServe keeps the deep learning model in memory and doesn't necessitate to reload it every time a new request arrives. It can also handle requests in parallels. Second, TorchServe can manage a pre-processing and a post-processing function that are defined in handlers. The main limitation of this work is that it does not take into account all the constraints related to production such as scaling, management of versions of the model, and verification of the stability of predictions over time.

This preliminary work is a first step for a research program on best practices related to the deployment of deep learning algorithms using Machine Learning Operations (MLOps), their advantages and disadvantages. It can serve as a basis for future comparisons with other types of architectures.

References

[1] Dupuy-Zini A, Audeh B, Gagneux-Brunon A, Bousquet C. Users' Reactions on Announced Vaccines against COVID-19 Before Marketing in France: Analysis of Twitter posts. medRxiv 2022.02.14.22268832; doi: https://doi.org/10.1101/2022.02.14.22268832

[2] Martin L, Muller B, Suárez PJO, Dupont Y, Romary L, de La Clergerie ÉV, et al. CamemBERT: a tasty French language model. arXiv preprint. 2019. arXiv:1911.03894. 2019.

[3] Devlin J, Chang MW, Lee K, Toutanova K. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint. 2018. arXiv:1810.04805. 2018.

Challenges of Trustable AI and Added-Value on Health
B. Séroussi et al. (Eds.)

doi:10.3233/SHTI220418

Robust Random Forest-Based All-Relevant Feature Ranks for Trustworthy AI

Bastian PFEIFER[a,1], Andreas HOLZINGER[a,b] and Michael G. SCHIMEK[a]
[a] *Institute for Medical Informatics Statistics and Documentations, Medical University of Graz, Austria*
[b] *Alberta Machine Intelligence Institute, Canada*

Abstract. Feature selection is a fundamental challenge in machine learning. For instance in bioinformatics, it is essential when one wishes to detect biomarkers. Tree-based methods are predominantly used for this purpose. In this paper, we study the stability of the feature selection methods BORUTA, VITA, and RRF (regularized random forest). In particular, we investigate the feature ranking instability of the associated stochastic algorithms. For stabilization of the feature ranks, we propose to compute consensus values from multiple feature selection runs, applying rank aggregation techniques. Our results show that these consolidated features are more accurate and robust, which helps to make practical machine learning applications more trustworthy.

Keywords. Feature Selection, Random Forest, Rank Aggregation, Trustworthy AI

1. Motivation

Feature selection is an important preprocessing step in many machine learning applications and has long been a fundamental challenge. In the biomedical field, feature selection is commonly used for data-driven biomarker discovery. Most common feature selection methods are based on the random forest (RF) classifier because it provides an interpretable mechanism for computing feature importance. Very important aspects of trustworthy AI are robustness and explainability [1]. A robust feature selector should (1) report on a constant set of relevant features when executed on exactly the same data set multiple times (stability), and (2) select the most relevant features for the modeling process of interest. In this work we analyzed the robustness of three feature selection methods which are widely used for data-driven biomarker discovery, namely the BORUTA algorithm [2], the VITA algorithm [3], and the RRF algorithm [4]. In essence, we are proposing to run the above mentioned algorithms multiple times and to consolidate the observed rank variations of importance scores through rank aggregation techniques. In simulation experiments on synthetic data we could show, that the RF consensus feature ranks obtained via rank aggregation can substantially improve the selection of the most important and best performing features.

[1] Corresponding Author, Bastian Pfeifer, Institute for Medical Informatics Statistics and Documentation, Medical University of Graz, Austria; E-mail: bastian.pfeifer@medunigraz.at.

2. Results

For evaluation we used the Madelon data sets from the the UCI Machine Learning Repository (https://archive.ics.uci.edu/ml/datasets/madelon). Madelon is synthetic and contains data points grouped in 32 clusters placed on the vertices of a five-dimensional hypercube randomly labeled +1 or -1. Madelon includes 20 relevant and 480 irrelevant features. The performance of the feature selection algorithms was analyzed while subsequentially down-sampling the data (15%, 25%, 50%, and 75%), leading to a stepwise reduction of the relevant signals. The feature selection algorithms were executed 20 times on exactly the same down-sampled data set. The rank aggregation techniques Borda (l2norm, mean, geometric mean, and median; [5]) and alternatively RRA [6] were applied for the consensus calculations. To evaluate the stability of the feature selectors *coverage* was used, defined as *Rs/Rt*, where *Rs* is the number of relevant features which are successfully selected and *Rt* is the total number of relevant features present in the data. The analysis of the Madelon data set indicates that VITA as well as BORUTA can be improved by consolidated consensus scores. When only 25% of the data are randomly sampled, the overall performance and robustness decreases. A detailed summary of this evaluation can be obtained from Table 1, where we also varied the number of trees within the RF. We observe that for BORUTA and VITA, the consolidated consensus ranks have a much higher coverage compared to the median coverage of the 20 feature selection runs. This observation indicates that the consensus calculation may have a higher impact on robustness than an increased number of trees. Even for a number of trees as high as 1000, there is a notable difference in performance between the worst (min coverage) and the best (max coverage) run. Furthermore, we identify BORUTA as the most accurate algorithm, especially when the number of trees is low and the signal is weak. However, it is computationally more demanding than VITA. Compared to BORUTA and VITA, RRF does not benefit much from the consensus calculations.

Table 1. Madelon data set. Coverage of the consensus ranks with varying numbers of trees.

FS Method	n.trees	min/median/max	l2norm	mean	geomean	median	RRA
BORUTA	100	0.32/0.50/0.74	0.79	0.79	0.84	0.84	0.79
	500	0.58/0.68/0.79	0.74	0.74	0.79	0.74	0.74
	1000	0.53/0.68/0.79	0.74	0.74	0.74	0.74	0.74
VITA	100	0.26/0.39/0.53	0.58	0.63	0.63	0.63	0.58
	500	0.42/0.58/0.68	0.53	0.63	0.63	0.58	0.63
	1000	0.52/0.68/0.79	0.74	0.74	0.74	0.74	0.74
RRF	100	0.37/0.47/0.58	0.26	0.37	0.47	0.47	0.47
	500	0.42/0.47/0.58	0.42	0.47	0.47	0.47	0.42
	1000	0.42/0.47/0.53	0.47	0.47	0.47	0.47	0.47

References

[1] Nogueira S, Sechidis K, Brown G. On the stability of feature selection algorithms. *Journal of Machine Learning Research*. 2017: 18(1): 6345-6398.
[2] Kursa MB, Rudnicki WR. Feature selection with the Boruta package. *J of Stat Soft*. 2010:36(11): 1–13.
[3] Janitza S, Celik E, Boulesteix AL. A computationally fast variable importance test for random forests for high-dimensional data. *Advances in Data Analysis and Classification*. 2018: 12(4): 885–915.
[4] Deng H, Runger G. Gene selection with guided regularized random forest. *Pattern Recognition*. 2013: 46(12): 3483-3489.
[5] Schimek MG, et al. Topklists: a comprehensive R package for statistical inference, stochastic aggregation, and visualization of multiple omics ranked lists. *Stat App in Gen Mol Bio*. 2015:14:311-6.
[6] Kolde R, Laur S, Adler P, and Vilo J. Robust rank aggregation for gene list integration and meta-analysis. *Bioinformatics*. 2012: 28(4): 573–580.

Challenges of Trustable AI and Added-Value on Health
B. Séroussi et al. (Eds.)

doi:10.3233/SHTI220419

Prediction of Acute Kidney Injury in the Intensive Care Unit: Preliminary Findings in a European Open Access Database

Michael FUJARSKI [a,1*], Christian PORSCHEN [b,*], Lucas PLAGWITZ [a]
Alexander BRENNER [a], Narges GHOREISHI [c] Patrick THORAL [d],
Harm-Jan DE GROOTH [d], Paul ELBERS [d], Raphael WEISS [b],
Melanie MEERSCH [b], Alexander ZARBOCK [b],
Thilo Caspar VON GROOTE [b,**] and Julian VARGHESE [a,**]
[*] *the authors contributed equally to this work and share the first authorship*
[**] *the authors contributed equally to this work and share the coordinating authorship*
[a] *Institute of Medical Informatics, University of Münster, Germany*
[b] *Department of Anesthesiology, Intensive Care and Pain Medicine, University Hospital Münster, Münster, Germany*
[c] *Federal Institute for Risk Assessment, Berlin, Germany*
[d] *Department of Intensive Care Medicine, Laboratory for Critical Care Computational Intelligence. Vrije Universiteit, Amsterdam, The Netherlands*

Abstract. Acute kidney injury (AKI) is a common complication in critically ill patients and is associated with long-term complications and an increased mortality. This work presents preliminary findings from the first freely available European intensive care database released by Amsterdam UMC. A machine learning (ML) model was developed to predict AKI in the intensive care unit 12 hours before the actual event. Main features of the model included medications and hemodynamic parameters. Our models perform with an accuracy of 81.8% on moderate to severe AKI and 79.8% on all AKI patients. Those results can compete with models reported in the literature and introduce an ML model for AKI based on European patient data.

Keywords. Acute Kidney Injury, AmsterdamUMCdb, ICU, Predictive Modeling

1. Introduction

Acute kidney injury (AKI) is a common complication in critically ill patients in the intensive care unit (ICU). AKI is diagnosed based on reductions of urinary output and elevations of serum creatinine, according to the KDIGO 2012 guidelines [1]. Since preventative measures effectively reduce the incidence of AKI the development of Machine Learning (ML) based models has been suggested for early identification [2]. Since the published models have mainly been developed on databases from the United States, this study, we developed such a model, based on European patient data to achieve an adequate generalizability [3].

[1] Corresponding Author, Michael Fujarski; E-mail: michael.fujarski@wwu.de.

2. Methods

The AmsterdamUMCdb contains data of 23,106 medical and surgical patients treated between 2003 and 2016 in the Amsterdam University Medical Centers. Available data includes routinely collected ICU data. A cohort of moderate or severe AKI were identified by an increase of their serum creatinine. A window of 48 hours was used to predict an AKI 12 hours into the future.

3. Results

The best performing model achieved an accuracy of 81.8% on severe AKI patients and 79.8% on all patients. See the supplements for detailed results and feature analyses [4].

4. Discussion

Using the AmsterdamUMCdb database, an ML based model for the prediction of AKI was developed and achieves a moderate accuracy. Frequently administered drugs that optimize hemodynamics as well as blood pressure lowering drugs correlated positively with the target variable. The limitations of the model include a missing validation on other databases and the identification of AKI solely based on serum creatinine.

5. Conclusion

The AmsterdamUMCdb offers the first open-access database of European ICU patient data. We present a novel ML model that accurately predicts severe AKI 12 hours prior to clinical occurrence. Future work includes a validation on other databases and implementation of other AKI criteria like reduced urine production.

References

[1] KDIGO Clinical Practice Guideline for Acute Kidney Injury. Official Journal of the International Society of Nephrology. 2012 Mar 1.

[2] Meersch M, Schmidt C, Hoffmeier A, Van Aken H, Wempe C, Gerss J, Zarbock A. Prevention of cardiac surgery-associated AKI by implementing the KDIGO guidelines in high risk patients identified by biomarkers: the PrevAKI randomized controlled trial. Intensive Care Med. 2017 Nov;43(11):1551-1561. doi: 10.1007/s00134-016-4670-3. Epub 2017 Jan 21. Erratum in: Intensive Care Med. 2017 Mar 7;: PMID: 28110412; PMCID: PMC5633630..

[3] Thoral PJ, Peppink JM, Driessen RH, Sijbrands EJG, Kompanje EJO, Kaplan L, Bailey H, Kesecioglu J, Cecconi M, Churpek M, Clermont G, van der Schaar M, Ercole A, Girbes ARJ, Elbers PWG; Amsterdam University Medical Centers Database (AmsterdamUMCdb) Collaborators and the SCCM/ESICM Joint Data Science Task Force. Sharing ICU Patient Data Responsibly Under the Society of Critical Care Medicine/European Society of Intensive Care Medicine Joint Data Science Collaboration: The Amsterdam University Medical Centers Database (AmsterdamUMCdb) Example. Crit Care Med. 2021 Jun 1;49(6):e563-e577. doi: 10.1097/CCM.0000000000004916. PMID: 33625129; PMCID: PMC8132908..

[4] Fujarski , Porschen. Sciebo University of Münster. [Online]. [cited 2022 03 31. Available from: https://uni-muenster.sciebo.de/s/ec3oTTfbfxwEWWT.

Challenges of Trustable AI and Added-Value on Health
B. Séroussi et al. (Eds.)

doi:10.3233/SHTI220420

Generation and Evaluation of Synthetic Data in a University Hospital Setting

Bayrem KAABACHI[a,b,1], Jérémie DESPRAZ[a] , Thierry MEURERS[c], Fabian PRASSER[c] and Jean Louis RAISARO[a]
[a] *Centre Hospitalier Universitaire Vaudois (CHUV), Lausanne, Switzerland*
[b] *Ecole Polytechnique Fédérale de Lausanne (EPFL), Switzerland*
[c] *Berlin Institute of Health @ Charité – Universitätsmedizin Berlin, Germany*

Abstract. In this study, we propose a unified evaluation framework for systematically assessing the utility-privacy trade-off of synthetic data generation (SDG) models. These SDG models are adapted to deal with longitudinal or tabular data stemming from electronic health records (EHR) containing both discrete and numeric features. Our evaluation framework considers different data sharing scenarios and attacker models.

Keywords. Synthetic Data, Privacy, Medical Research, Generative Adversarial Networks

1. Introduction

Synthetic Data Generation (SDG) models, and especially Generative Adversarial Networks (GANs) [1] are innovative and rapidly evolving tools spanning a wide range of fields that can produce new samples from a reference dataset with similar statistical properties. This approach is increasingly considered for the generation of synthetic electronic health records (EHRs) [2]. However, there is a lack of a standardized evaluation of these tools, both in terms of utility and privacy.

Our main contributions are: (i) improvement of the SynTEG framework, [3], state-of-the-art GAN-based SDG model for longitudinal EHR data, (ii) review of the most used utility metrics for both tabular and longitudinal synthetic medical data (iii) adaptation of a privacy-evaluation [4] with regards to hospital-related data sharing use cases and the ability to model various degrees of adversarial background knowledge, and (iv) development of a unified, practical and modular framework that encompasses both utility and privacy metrics for a systematic evaluation of synthetic medical data.

2. Methods

Generative models: We selected the following two state-of-the art GAN-based models: CTGAN [5] for tabular and SynTEG [3] for longitudinal data. We adapted SynTEG to handle longitudinal EHR data that do not contain those diagnosis codes, and that include continuous variables such as vital parameters. We propose a utility evaluation module

[1] Corresponding Author, Bayrem KAABACHI; E-mail: bayrem.kaabachi@gmail.com.

that encompasses the most used utility metrics described in the literature that we organized in six categories. This module characterizes the extent to which our models capture correlations, distributions and temporal patterns of the original dataset. We evaluated the capacity of the models to protect against privacy attacks by adapting and improving the synthetic data privacy evaluation approach developed by Stadler et al. [4]. We enabled the framework to deal with attackers having various partial prior knowledge. We are thus able to measure the risk of membership inference and attribute inference attacks for each patient record. These modules are combined in a unified framework that encompasses both utility and privacy evaluation [2].

3. Results

Results are based on a public dataset provided by the Texas Health Department [6]. The following figure shows a web-based interactive dashboard through which the user can assess the results of multivariate and univariate statistical analysis and visually compare the results between the real and synthetic dataset. We compute the privacy gain PG [4]. A high privacy gain indicates that the probability of re-identification is greatly reduced when disclosing the synthetic data rather than the original dataset. An example report is provided in the Git repository.

4. Discussion and Conclusion

We proposed a unified benchmark to understand the potential gains and risks of using synthetic data. Generative adversarial networks could be a way to fast-forward the development of new AI-based clinical decision support systems or to easily access look-alike data for training and education purposes. Yet, we strongly believe that a formal and common evaluation metric is crucial.

References

[1] Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, et al. Generative adversarial nets. In: Advances in neural information processing systems; 2014. p. 2672-80.
[2] El Emam K, Mosquera L, Jonker E, Sood H. Evaluating the utility of synthetic COVID-19 case data. JAMIA Open. 2021 Jan;4(ooab012). Available from: https://doi.org/10.1093/jamiaopen/ooab012.
[3] Zhang Z, Yan C, Lasko TA, Sun J, Malin BA. SynTEG: a framework for temporal structured electronic health data simulation. Journal of the American Medical Informatics Association. 2020 Nov;(ocaa262)
[4] Stadler T, Oprisanu B, Troncoso C. Synthetic Data – Anonymisation Groundhog Day. In: 31st USENIX Security Symposium (USENIX Security 22). Boston, MA: USENIX Association; 2022.
[5] Xu L, Skoularidou M, Cuesta-Infante A, Veeramachaneni K. In: Modeling Tabular Data Using Conditional GAN.Red Hook, NY, USA: Curran Associates Inc.; 2019
[6] TDSHS. Hospital Discharge Data Public Use Data File.

[2] The utility-privacy evaluation framework package can be accessed at https://gitlab.itrcs3-app.intranet.chuv/bkaabachi/evaluation-framework

Challenges of Trustable AI and Added-Value on Health
B. Séroussi et al. (Eds.)

doi:10.3233/SHTI220421

Conversational Agent to Address COVID-19 Infodemic: A Design-Based Research Approach

Youness ZIDOUN[a,1], Sreelekshmi KALADHARA[a], Yehia KOTP[b], Leigh POWELL[a], Radwa NOUR[a], Hanan AL SUWAIDI[c] and Nabil ZARY[a]
[a] *IeHPE, MBRU, Dubai, UAE*
[b] *Smart Services, MBRU, Dubai, UAE*
[c] *College of Medicine, MBRU, Dubai, UAE*

Abstract. Since the beginning of the year 2020, we have been suffering from the COVID-19 pandemic and are daily exposed to misinformation, leading to myths around vaccination and COVID-19. This study focuses on creating and distributing a Conversational Agent (CA), named VWise, for a health intervention using Design-Based Research (DBR), to help profile, guide, and inform the public about COVID-19 and COVID-19 vaccination in the EMRO (Eastern Mediterranean Region of Operations) region.

Keywords. Conversational Agent, NLU, Design-Based Research, COVID-19 vaccination.

1. Introduction

Vaccine hesitancy is a global issue, described as the refusal or delay of vaccine uptake notwithstanding the availability of service [1]. To impact the public's behavior, health communication approaches should activate community participation and provide avenues for individuals to receive and communicate their needs [1]. This study aims on designing a Conversational Agent (CA) using Design-Based Research to help profile and inform the public about COVID-19 and COVID-19 vaccination in the EMR region.

2. Methods and Results

The purpose of the Design-Based Research (DBR) is to generate new theories and frameworks for conceptualizing learning, instruction, design processes, and educational reform [3]. This study focuses on the first iteration of the Vwise design, using the multiple cycles of the DBR approach - Design, Analyze and Review. After completing the first iteration of the design cycle with the initial NLU model derived from the WHO FAQ (Frequently Asked Questions), an improved model will be obtained to inform the next phase, using Conversation Driven Development (CDD) process.

[1] Corresponding Author, Youness ZIDOUN; E-mail: youness.zidoun@mbru.ac.ae

VWise integration architecture as per figure 1, has a training model, distribution channel, separate database to store conversations, finally the collected data analyzed, to draw conclusions out of it. The entire navigation flow is set up in two separate Virtual Machines, one hosts the web pages, and the other hosts the CA GUI which is the Rasa X instance.

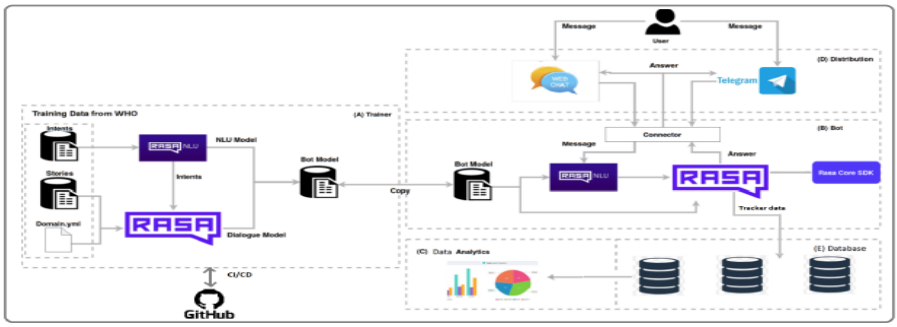


Figure 1. The VWise integration architecture

3. Discussion and Conclusions

Even though using DBR is quite appealing in linking development and research as closely as possible, evidence suggests that significant and transferable results can only be achieved if numerous iterations are executed. Owing to the project's timeline, we propose that the necessary phases of DBR will be followed even in this short time frame with limited number of iterations, to generate acceptable and practical recommendations for designing a contextualized CA, which addresses vaccine hesitancy. The current stage consists of the first iteration of the design, having covered the two cycles of analysis and design. The evaluation and reflection phase will take place right after the end of the last CDD process. As a pragmatic methodology, DBR can represent a holistic approach to inform such innovative projects considering its iterative principles.

4. Acknowledgement

This work is supported in part by the Al Jalila Foundation and in addition funded by a Pfizer Independent Medical Education Grant (Agreement No. 67504787).

References

[1] '1_Report_WORKING_GROUP_vaccine_hesitancy_final.pdf'.Accessed:Jan.16,2022.[Online]. Available:https://www.who.int/immunization/sage/meetings/2014/october/1_Report_WORKING_GROUP_vaccine_hesitancy_final.pdf

[2] Walwema J. The WHO Health Alert: Communicating a Global Pandemic with WhatsApp, Journal of Business and Technical Communication. 2021 Jan; 35(1), 35–40, doi: 10.1177/1050651920958507

[3] Johnson C, et al. Using design-based research to develop meaningful online discussions in undergraduate field experience courses. International Review of Research in Open and Distributed Learning. 2017; 18(6). Retrieved from http://www.irrodl.org/index.php/irrodl/article/view/2901/4373

Challenges of Trustable AI and Added-Value on Health
B. Séroussi et al. (Eds.)

doi:10.3233/SHTI220422

Prediction of Hospitalization Using Machine Learning for Emergency Department Patients

Georgios FERETZAKIS[a,b,c,1], Aikaterini SAKAGIANNI[d], Evangelos LOUPELIS[c], Dimitris KALLES[a], Vasileios PANTERIS[e], Lazaros TZELVES[f], Rea CHATZIKYRIAKOU[g], Nikolaos TRAKAS[h], Stavroula KOLOKYTHA[i], Polyxeni BATIANI[i], Zoi RAKOPOULOU[j], Aikaterini TIKA[j], Stavroula PETROPOULOU[c], Ilias DALAINAS[j] and Vasileios KALDIS[i]

[a] *School of Science and Technology, Hellenic Open University, Patras, Greece*
[b] *Department of Quality Control, Research and Continuing Education, Sismanogleio General Hospital, Marousi, Greece*
[c] *IT department, Sismanogleio General Hospital, Marousi, Greece*
[d] *Intensive Care Unit, Sismanogleio General Hospital, Marousi, Greece*
[e] *Gastroenterology Department, Sismanogleio General Hospital, Marousi, Greece*
[f] *Second Department of Urology, National and Kapodistrian University of Athens, Sismanogleio General Hospital, Marousi, Greece*
[g] *Hematology Laboratory, Sismanogleio General Hospital, Marousi, Greece*
[h] *Biochemistry Department, Sismanogleio General Hospital, Marousi, Greece*
[i] *Emergency Department, Sismanogleio General Hospital, Marousi, Greece*
[j] *Administration, Sismanogleio General Hospital, Marousi, Greece*

Abstract. The objective of this study was to evaluate the predictive capability of five machine learning models regarding the admission or discharge of emergency department patients. A Random Forest classifier outperformed other models with respect to the area under the receiver operating characteristic curve (AUC ROC).

Keywords. machine learning, emergency department, patient admission

1. Introduction and Background

One of the greatest challenges that most Emergency Departments (ED) face daily is the surge of patient volume and the limited medical resources capacity. Fast recognition of emergency problems requiring direct admission to the hospital is of utmost importance to ensure an efficient triage workflow [1, 2]. In the present study, by using routine ED data and laboratory exams, we evaluated five ML algorithms to promptly forecast patient hospital admission and discharge, aiming at reducing the increased workload of emergency specialists, reducing emergency department congestion, and improving patient care.

[1] Corresponding Author, Georgios FERETZAKIS, PhD, Greece; E-mail: georgios.feretzakis@ac.eap.gr

2. Methods and Materials

This observational retrospective research was conducted in the ED of a public tertiary care hospital in Greece and has been approved by the Institutional Review Board of Sismanogleio General Hospital. The experimental tasks are carried out using the caret package in R [3]. During a one-year period from January until December 2020, there were recorded 13,991 ED visits from patients who had undergone routine laboratory testing. Apart from demographics (Age, Gender), we also included in our analysis biochemical markers and coagulation tests in patients visiting the ED alongside the ED outcome (admission or discharge, i.e. we deal with a binary classification problem). Multiple Imputations by Chained Equations (MICE) method [4] for missing data imputation was implemented and used with the R programming language. We applied and evaluated the following classification models: linear discriminant analysis (method: lda), Recursive Partitioning and Regression Trees (method: rpart), support vector machines (method: svmRadial), k – nearest neighbor (method: knn), and random forests (method: rf)[5].

3. Results

Among the different classifiers that were evaluated through 10-fold cross-validation, a random forest model outperformed other models with respect to AUC ROC. The performance metrics when we applied the five classifiers in the whole data set are presented in the following table (Table 1).

Table 1. Performance metrics of the five classifiers in the whole data set

Metric/Method	lda	rpart	knn	svmRadial	rf
AUC ROC	0.7569	0.6995	0.6832	0.7788	0.7813

4. Discussion and Conclusions

In this study, we evaluated five classifiers to predict admission or discharge of emergency department patients. A random forest model outperformed other models with respect to AUC ROC. The proposed ML technique would act as an assisting tool for clinicians in a way that its output summarizes part of the attributes that the clinicians take into consideration to decide about hospital admission, especially in crowded emergency departments where the available human resources are often limited.

References

[1] Sun BC, Hsia RY, Weiss RE, et al. Effect of emergency department crowding on outcomes of admitted patients. Ann Emerg Med. 2013 Jun;61(6):605-611.e6.

[2] Feretzakis G, Sakagianni A, Loupelis E, et al. Predicting Hospital Admission for Emergency Department Patients: A Machine Learning Approach. Stud Health Technol Inform. 2022 Jan 14;289:297-300.

[3] Kuhn M. Building Predictive Models in R Using the caret Package. Journal of Statistical Software. 2008; 28(5), 1–26.

[4] Van Buuren, S., & Groothuis-Oudshoorn, K. mice: Multivariate Imputation by Chained Equations in R. Journal of Statistical Software, 2011;45(3):1–67.

[5] Breiman, L. (2001) Random Forests. Machine Learning, 45:5-32.

Challenges of Trustable AI and Added-Value on Health
B. Séroussi et al. (Eds.)

doi:10.3233/SHTI220423

Learning Bayesian Networks for the Prediction of Unfavorable Health Events in Nursing Homes

Clara CHARON[a,b,1], Pierre-Henri WUILLEMIN[b] and Joël BELMIN[c,d]
[a] *Teranga Software, Paris, France*
[b] *LIP6 (UMR 7606), Sorbonne Université, Paris, France*
[c] *AP-HP, Hôpital Charles-Foix, Ivry-sur-Seine, France*
[d] *LIMICS (UMR 1142), Sorbonne Université, Paris, France*

Abstract. This study[2] proposes the use of Bayesian Networks for the prediction of unfavorable health events, and more especially pressure ulcers, in nursing homes. From a database of electronic medical records, we learn an explainable and relevant classifier, which performs better than the scores currently used in nursing homes.

Keywords. Machine Learning, Bayesian Networks, Nursing Homes, Explainability

1. Introduction and Methods

With the development of new computer technologies in the healthcare field, a growing number of healthcare institutions use information systems that gather administrative and medical data on patients. NETSoins, edited by Teranga Software, is one of these systems and is used by more than 3 000 nursing homes (NH). The goal is to use anonymous data from this NH residents record software to develop algorithms able to predict several adverse health events that are potentially preventable by appropriate health interventions. As a first step, the prediction is focused on the occurrence of a resident's first pressure ulcer (PU). A PU is a lesion of the skin related to prolonged compression between a bone and the support on which the patient rests. This condition is painful and a source of infection, and it significantly impairs quality of life [1]. PU are however highly preventable with a particular and multidisciplinary approach. Currently, NH staff can use simple clinical tools to identify the residents at risk to develop PU, like the Norton and the Braden scales, but they overlook important risk factors [2]. To develop a better preventive system, we particularly focused on artificial intelligence methods with strong explainability such as Bayesian Networks (BN). A BN is a probabilistic graphical model that represents a set of variables and their conditional dependencies via a directed acyclic graph [3]. BN are often considered as a good compromise between accuracy and explainability. Here, we make the choice to focus on bringing knowledge from a medical point of view (like new risk factors). This also allows the users, carers and medical staff, to have confidence because they can understand the algorithm's decision. In this process,

[1] Corresponding Author, Clara Charon, LIP6, 4 Place Jussieu, 75005 Paris; E-mail: clara.charon@lip6.fr.
[2] This research is funded by Teranga Software and the ANRT CIFRE Grant n°2019/1519 for CC.

we have retrieved information from medical records of more than 100 000 residents free of PU at admission in the facility. Variables with potential interest were determined with expert geriatricians and kept in the analysis based on the completion rate. A training dataset was then set up with the parameters of the residents and two classes were defined: experience the PU or not. An important issue is to take into account the transformation of an event database into data that can be used by statistical learning, while keeping a medical meaning. To elaborate a BN classifier, we learn the structure of the network from estimates on the training base, and then compute the probability of belonging to a class knowing the observations of each variable on an individual. For the evaluation, we compute the F-score. This is the harmonic mean of the precision and sensitivity scores and we want to maximize it. Three distinct datasets with different temporalities objectives are created: a prediction 1 month before the PU episode, 2 and 3 months before.

2. Results

Table 1. Summary tables of the different F-scores according to the methods and prediction timings

F-Score	1-month	2-months	3-months
BN Classifier	0.70	0.69	0,67
Random Forest	0.72	0.69	0.70
Braden Scale	0.32	-	-
Norton Scale	0.29	-	-

The results obtained are available in Table 1. We notice for pyAgrum's BN Classifier [4], that the more the timeline is far away, the more difficult it is to correctly predict the PU. However, the results do not differ much and the benefit for the resident to have an earlier preventive protocol is important. The F-Score of BN Classifier has been compared with that of Random Forest and we found the predictive power was similar. We also calculated the F-score of the Norton and Braden scales and our method has much better results than the methods currently used in NH.

3. Discussion and Conclusions

In this paper, we propose a classification based on BN in a predictive medical context that proves to be efficient in terms of results. It would also be calculated automatically, which is a significant time saving for the caregiver. Many improvements are still possible, in particular in the exploitation of the time series. The intent is that the approach will be put into practice to generate alerts to bring high-risk situations to encourage to implement preventive interventions for the targeted events.

References

[1] Belmin J, et al. Gériatrie: pour le praticien. Elsevier Health Sciences; 2016 Sep 7..
[2] Kaewprag P, et al. Predictive modeling for pressure ulcers from intensive care unit electronic health records. AMIA Summits on Translational Science Proceedings. 2015;2015:82.
[3] Pearl J. Causality. Cambridge university press; 2009 Sep 14.
[4] Ducamp G, Gonzales C, Wuillemin PH. aGrUM/pyAgrum: a toolbox to build models and algorithms for Probabilistic Graphical Models in Python. InInternational Conference on Probabilistic Graphical Models 2020 Feb 2. PMLR.

Challenges of Trustable AI and Added-Value on Health
B. Séroussi et al. (Eds.)

doi:10.3233/SHTI220424

Extraction of Tumor Response Criteria in Semi-Structured Imaging Report

Valentin POHYER[a], David BAUDOIN[a,b], Laure FOURNIER[a,c,d] and Bastien RANCE[a,b,d,e]
[a] *Hôpital Européen Georges Pompidou, AP-HP, Paris, France*
[b] *INSERM, UMRS 1138, Centre de Recherche des Cordeliers, Université Sorbonne-Paris Cité, Paris, France*
[c] *INSERM, PARCC, Paris, France*
[d] *Université de Paris, Paris, France*
[e] *INRIA, France*

Abstract. In this study, we extracted information from 6,376 french CT scan semi-structured text reports evaluating the cancer treatment response using the RECIST methodology. We evaluated the performance against manual annotation of 100 reports and measured the evolution of the presence of information over time. The results show high performances of the extraction as well as trends.

Keywords. information retrieval, natural language processing, imaging, cancer

1. Introduction

Biomedical research used to rely mostly on structured data, however an important part of biomedical information is kept solely in the form of free text reports. Traditional approaches including the identification of regular expressions (regex) have been successfully used in numerous applications [1]. The Response Evaluation Criteria in Solid Tumors (RECIST 1.1) [2] is a methodology used in clinical research and clinical care to evaluate the efficacy of cancer treatments in solid tumors. The radiologists evaluate the evolution of tumors along three axes: target lesions (for which quantitative measurement are performed), non-target lesions, and the appearance of new lesions. An overall conclusion is computed depending on the combination of criteria of the three axes: complete response, partial response, stable disease, progressive disease or not-assessable.

2. Methods

We queried the clinical data warehouse at the European Hospital Georges Pompidou, located in Paris, France, to identify all semi-structured CT reports between January 1st, 2013 and December 31st, 2021. 6,376 reports were identified. We built a set of 243 regex to identify 20 key items using a trial and error process. We applied the regex on the corpus and analyzed errors. We leveraged Py-Rex, a Python tool developed in-house to extract information using the set of regex (code available at:

https://github.com/Vitalizful/Py-Rex). To evaluate the performance, we annotated manually a random sample of 100 documents for three items (baseline date, evolution of the size of target tumors and global RECIST response) and affected the extracted entities to either true positive, true negative, false positive or false negative. We used precision, recall and F1-score to describe the performance of the extractor (see Table 1). To evaluate the stability and quality of the extraction over the study period, we calculated the ratio of the number of data identified over the number of documents in a given period for every trimester.

3. Results and Discussion

Figure 1 shows the evolution of the ratio over the 8 years of the study. Our system captured increasingly more baseline data overtime (increasing from less than 20% of documents to over 60%). Conversely, the identification and extraction of the evolution of the size of target lesion gradually dropped from 90% to 60% over the period of study. Finally, the global RECIST conclusion remains stable overtime at almost 100% ratio of extraction.

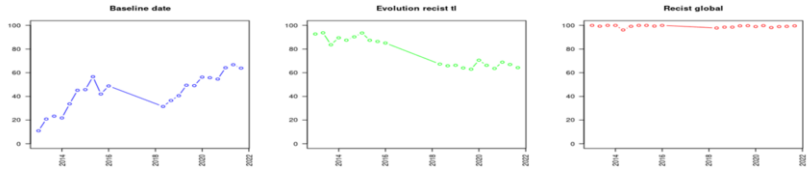


Figure 1. Stability of the extraction over time. The y-axis describes the ratio of documents in which an item is identified.

Table 1. Performance of the extraction

	Precision	Recall	F1-Score
Baseline date	97.8%	95.7%	96.8%
RECIST evolution	97.3%	100%	98.6%
Global conclusion	99%	100%	99.5%

Exchanges between the data scientists and radiologist are required to fully take advantage of the semi-structured documents. The overall performance of our method for the items of interest was very high (over 95%). The application of regex on semi-structured documents considerably simplifies the extraction process. We detected three types of stability profiles: increase, decrease and stability of items. Overall the evolution of the template of the semi-structured documents could be seen as an evolutionary process. The evolution of the structuring of an item is the product of selective pressure, mutation and drift. Our study used solely regex to identify and extract items of interest; however state of the art methods for entity recognition rely on machine learning methods, including neural networks (especially transformers architectures). These methods require large sets of manual expert annotations and often need to be retrained if new elements were to be added. In contrast, regex can be easily developed by radiologists and data scientists together.

References

[1] AAlAbdulsalam AK, et al. Automated Extraction and Classification of Cancer Stage Mentions from Unstructured Text Fields in a Central Cancer Registry. AMIA Jt. Summits Transl. Sci. 2018:16–25.

[2] Eisenhauer EA, et al. New response evaluation criteria in solid tumours: revised RECIST guideline (version 1.1), Eur. J. Cancer Oxf. Engl. 1990. 45(2009):228–247. doi:10.1016/j.ejca.2008.10.026.

Challenges of Trustable AI and Added-Value on Health
B. Séroussi et al. (Eds.)

doi:10.3233/SHTI220425

Reliability of Drug-Drug Interaction Measurement on Real-Word Data: The ReMIAMes Project

Catherine DUCLOS[a,1], Nicolas GRIFFON[a,b], Christel DANIEL[a,b], Guillaume BOUZILLÉ[c], Denis DELAMARRE[c], Stefan DARMONI[a,d], Laurent TOUBIANA[a] and Julien GROSJEAN[a,d]

[a] *Sorbonne Université, Université Sorbonne Paris Nord, INSERM, Laboratoire d'Informatique Médicale et d'Ingénierie des connaissances en e-Santé, LIMICS, F-75006 Paris, France*

b Innovation and Data, IT Department, AP-HP, Paris, France

[c]*Inserm, Laboratoire Traitement du Signal et de l'Image - UMR 1099, Centre Hospitalier Universitaire de Rennes, Université de Rennes 1, Rennes, France*

[d]*Department of Digital Health, Rouen University Hospital, France*

Abstract. The ReMIAMes project proposes a methodological framework to provide a reliable and reproducible measurement of the frequency of drug-drug interactions (DDI) when performed on real-world data. This framework relies on (i) a fine-grained and contextualized definition of DDIs, (ii) a shared minimum information model to select the appropriate data for the correct interpretation of potential DDIs, (iii) an ontology-based inference module able to handle missing data to classify prescription lines with potential DDIs, (iv) a report generator giving the value of the measurement and explanations when potential false positive are detected due to a lack of available data. All the tools developed are intended to be publicly shared under open license.

Keywords. Methodology, Quality, Data warehouse, drug drug interaction

1. Introduction

Drug-drug interactions (DDI) are a major source of adverse events and care consumption (1) that could be avoided if they were taken into account at the time of prescribing. Care data generated during hospitalization are now integrated in clinical data warehouses (CDW). Detection of DDIs present in CDWs is therefore a mean to control the quality of prescriptions and their pharmaceutical review process. To provide reliable and reproducible results, observational studies on real-world data have to rely on robust methodological basis (2). Many factors related to data source, DDI definition, drug attributes or patient should be taken into account to correctly analyze DDIs (3).

The objective of the project ReMIAMes is to propose a methodological framework to provide a reliable and reproducible measurement of DDIs frequency when performed on different CDWs.

[1] Corresponding Author, Catherine Duclos, LIMICS, Université Sorbonne Paris Nord, 74 rue Marcel Cachin, Bobigny, France; E-mail: catherine.duclos@inserm.fr

2. Methods

The French Thesaurus of DDIs (4) is the reference material for DDIs. These DDIs are analyzed to identify contextual attributes that may restrict their scope. It results in a conceptual model of DDI and requirements on minimum dataset to be provided by CDWs that are translated into a common model in a standard formalism (FHIR). Three University Hospital (APHP, Rouen, Rennes) participate. They have audited their data (drug administration data, lab data, clinical data) to identify their availability, their formalism. Several cycles of process of extraction, transfer, loading (ETL) are performed depending of the difficulty of providing certain requested data and according to a pattern of increasing complexity. A DDI ontology based on OWL formalism is modeled. Data received from CDWs are considered as instances and are used to classify pairs of prescription lines as potential DDI. According to the availability of information to decide, detected DDIs are classified as true positives, or potential false positives. Several rounds of data analysis are performed to identify the causes of false positives and let data providers to improve their ETL to reduce these false positive.

3. Results

The project will result in recommendations for ETL, open data to populate the ontology, ontology of DDIs and algorithms to analyze the data. A platform will integrate all these components.

4. Discussion Conclusion

The project ReMIAMes project is part of the production of methodological frameworks for the analysis of CDW data. It aims to improve the relevance and the quality of CDW data and to raise awareness of their impact in order to make analyses more reliable.

Acknowledgments

This study was granted by ReMIAMes (Ressources et méthodes libres pour la détection des interactions médicamenteuses dans les entrepôts de données de santé) ANR-19-CE19-0025-01

References

[1] Dechanont S, Maphanta S, Butthum B, Kongkaew C. Hospital admissions/visits associated with drug-drug interactions: a systematic review and meta-analysis. Pharmacoepidemiol Drug Saf. mai 2014;23(5):489-97.

[2] Agency for Healthcare Research and Quality. Developing a Protocol for Observational Comparative Effectiveness Research: A User's Guide.

[3] Horn J, Ueng S. The Effect of Patient-Specific Drug-Drug Interaction Alerting on the Frequency of Alerts: A Pilot Study. Ann Pharmacother. nov 2019;53(11):1087-92.

[4] Agence nationale de sécurité du médicament et des produits de santé. Thésaurus des interactions médicamenteuses [Internet]. Disponible sur: https://ansm.sante.fr/documents/reference/thesaurus-des-interactions-medicamenteuses-1

Challenges of Trustable AI and Added-Value on Health
B. Séroussi et al. (Eds.)

doi:10.3233/SHTI220426

Nursing Documentation and Activities when Using a Clinical Decision Support System for Sepsis and Clinical Deterioration Management

Luciana Schleder GONÇALVES[a,1], Maria Luiza Medeiros AMARO[b], Fernanda Karolinne SCHAMNE[b], Ingrid Marcela Pinto GARIBA[c]
[a] *Nursing Department, Federal University of Parana, Brazil*
[b] *Nursing Graduate Program, Federal University of Paraná, Brazil*
[c] *Caring Practices Professional Master Program, Federal University of Paraná, Brazil*

Abstract. The role of nurses in the management of sepsis and clinical deterioration, and the potential use of Clinical Decision Support Systems to identify patients' undesired outcomes are well recognized. In order to verify the association of such a system adoption in the compliance of a hospital sepsis protocol on the nurses' workflow and documentation, a three-stage study was conducted. Main findings show that there is no statistically significant difference between the sepsis protocol compliance data from before and after the system implementation; that nursing caring-related activities are a priority over documentation and time spent using the system; and that it was possible to validate a proposal of nursing documentation for sepsis and clinical deterioration. This contributes to the improvement of nurses' awareness of their engagement in nursing informatics issues in the era of big data.

Keywords. Nursing informatics, clinical decision support system, sepsis.

1. Introduction

Sepsis is a significant healthcare issue in Brazil, with death numbers ranging from 982,294 reported in 2002 to 1,133,761 as reported in 2010 census. Nurses' role in the assessment of first sepsis signs and symptoms is already well recognized [1]. The sepsis protocol of a hospital that co-hosted this study was used for a Clinical Decision Support Systems (CDSS) framework, based on the Modified Early Warning Score – MEWS [2]. This study had the objective to investigate the potential association of the adoption of such system in the compliance of a Brazilian oncology hospital sepsis protocol and on the nurses' workflow and documentation.

[1] Corresponding Author, Av. Lothario Meissner, 623 Bloco Diático II, Nursing Department. Jardim Botânico, Curitiba-PR-Brazil 80210-170 E-mail: lsgk@ufpr.br

2. Methods

Developed during 2018 and 2019, the study comprehends three stages: a pre-exploratory, exploratory and comparative analysis of septic patient registries before and after the CDSS implementation; an observational study of nursing activities and a validation study of a nursing documentation proposal. It was approved by the hospital IRB.

3. Results

Sepsis was more prevalent among male patients of 55 years, on average. Consciousness levels ($p = 0.0001$) and systolic blood pressure ($p = 0.0004$) registries were more prevalent. A significant decrease in the number of deaths registries ($p = 0.0134$), an increase in the length of stay ($p = 0.0001$) and an increase in the mean of vital signs registries per patient ($p > 0.0001$) among registries from the post implementation period.

In a four-hour observation period, nursing technicians spent an average of 1h22 min., nurses 55min. on direct patient care; 54 min. on indirect patient care, nurses 1h24 min.; 46 min. on personal care, nurses 29 min.; 28 min. on nursing records, nurses 44min.; 20 min. on urgent care, nurses 22 min.; and 10 min. with the CDSS, nurses 23min.

Finally, out of the 83 possible nursing terms related for septic patients care, 51 presented a Coefficient of Validation Index greater than 0.8 [2].

4. Discussion

Nurses must be aware of the variety of health data that can be converted into information, knowledge and wisdom and of their ability to use them to predict adverse events. This should encourage their engagement in nursing informatics issues, specially in the era of big data [3]. Limitations mainly refers to the analysis under the learning curve of the CDSS adoption in the clinical setting. Also, to the characteristics of oncology patients, wthat may configure a bias considering sepsis and clinical deterioration protcols.

5. Conclusions

The evaluation of CDSS adoption in clinical settings drives science development in the field of health informatics. Its implementation strategy and adjustment to the nursing staff workflow might influence in protocols compliance. Nursing documentation standards improves nursing wisdom visibility, a key issue in the era of big data.

References

[1] Gonçalves LS; Amaro MLM; Romero ALM; Schamne FK; Fressatto JL; Bezerra CW. Implementation of an Artificial Intelligence Algorithm for sepsis detection. Rev Bras Enferm. 2020;73(3):e20180421. doi: http://dx.doi.org/10.1590/0034-7167-2018-0421

[2] Gonçalves LS, Andreatta D, Schamne FK. Cross-Mapping Study of Nursing Practice Terms from a Brazilian Hospital Database. Stud Health Technol Inform. 2020 Jun 16;270:1233-1234. doi: 10.3233/SHTI200378. PMID: 32570595.

[3] Macieira, T. G. R. et al. Evidence of progress in making nursing practice visible using standardized nursing data: A systematic review, AMIA Annu. Symp. Proc., v. 2017, p. 1205-14, 2017.

Challenges of Trustable AI and Added-Value on Health
B. Séroussi et al. (Eds.)

doi:10.3233/SHTI220427

Tracking Temporal Clusters from Patient Networks

Judith LAMBERT [a,b,1], Anne-Louise LEUTENEGGER[e], Anne-Sophie JANNOT[a,c] and Anaïs BAUDOT[b,d]
[a]*INSERM, CRC, Team 22, UMR1138, Paris, France* ; [b]*INSERM, CNRS, Aix Marseille Univ, MMG, UMR1251, Marseille, France* ; [c]*Department of Statistics, Medical Informatic and Public Health, HEGP, AP-HP* ; [d]*Barcelona Supercomputing Center, Barcelona, Spain* ; [e]*INSERM, Université de Paris, NeuroDiderot, UMR1141, Paris, France*

Abstract. Creating homogeneous groups (clusters) of patients from medico-administrative databases provides a better understanding of health determinants. But in these databases, patients have truncated care pathways. We developed an approach based on patient networks to construct care trajectories from such truncated data. We tested this approach on antithrombotic treatments prescribed from 2008 to 2018 contained in the échantillon généraliste des bénéficiaires (EGB). We constructed a patient network for each patients' age (years from birth). We then applied the Markov clustering algorithm in each network. The care trajectories were finally constructed by matching clusters identified in two consecutive networks. We calculated the silhouette score to assess the performance of this network approach compared to three existing approaches. We identified 12 care trajectories that we were able to associate with pathologies. The best silhouette score was obtained for the network approach. Our approach allowed to highlight care trajectories taking into account the longitudinal, multidimensional and truncated nature of data from medico-administrative databases.

Keywords. Longitudinal clustering, Patient networks, Care trajectories

1. Introduction

Finding homogeneous groups (clusters) of patients in medico-administrative databases helps to better understand health determinants and improve patient prognosis. But the longitudinal and multidimensional data of each patient are available at different times of their life in these databases, thus forming truncated care pathways. Clustering such data is therefore challenging. There are three types of approaches for clustering longitudinal data[1]. But the number of clusters must be specified, missing values are not handled or must be imputed and only one longitudinal factor at a time can be considered.

We developed an approach to construct care trajectories based on patient networks. Patient networks have the advantage of handling heterogeneous and missing data, preserving patient privacy and not requiring to prespecify the number of clusters[2]. As a use case, we analyzed drug prescriptions contained in the échantillon généraliste des bénéficiares (EGB).

[1] Corresponding Author, Judith LAMBERT; E-mail: judith.lambert@crc.jussieu.fr

2. Material and Methods

2.1 Echantillon généraliste des bénéficiaires (EGB)

The EGB is a random sample from the French health insurance database recording healthcare reimbursements from approximately 660,000 individuals followed for 11 years. We extracted all prescriptions for antithrombotic agents from 2008 to 2018 among patients aged 60 to 70 resulting in a sample of 30,111 patients.

2.2 A patient clustering approach based on patient networks

A patient network is a graph defined with nodes representing patients and edges representing the similarity between patient nodes. We constructed a patient network for each patients' age by computing the similarity using the Cosine similarity[3]. We applied the Markov clustering algorithm[4] on each patient network to identify patient clusters. We then tracked these clusters based on the number of patients they have in common between each consecutive age.

2.3 Competitive models

We compared the above approach with three existing approaches. As these approaches imply to set a priori the number of clusters, we chose the number of trajectories identified with the network approach. Their clustering performance was compared by calculating the silhouette score.

3. Results

We identified with the network approach 12 care trajectories characterized by a specific predominant drug and composed of 100 patients at least. The best silhouette score was obtained with the network approach (0.44).

4. Discussion

Among the 12 care trajectories identified with the network approach, we were able to associate three of them with stroke, infarction and arrhythmia. The clustering performance was better with the network approach compared to the three existing approaches. The use of patient networks therefore allowed to take into account the longitudinal, multidimensional and truncated nature of data.

References

[1] Liao TW. Clustering of time series data—a survey. Pattern recognition, 2005, vol. 38, no 11, p. 1857-1874.

[2] Pai S, Bader GD. Patient similarity networks for precision medicine. Journal of molecular biology, 2018, vol. 430, no 18, p. 2924-2938. Singhal A. Modern information retrieval: A brief overview. IEEE Data Eng. Bull., 2001, vol. 24, no 4, p. 35-43.

Section II

Health Information Systems

This page intentionally left blank

Challenges of Trustable AI and Added-Value on Health
B. Séroussi et al. (Eds.)

doi:10.3233/SHTI220429

Lack of Human Resources Leads to Breaches in Information Management Processes

Heli IKONEN[a,1], Virpi JYLHÄ[a], Hanna KUUSISTO[ab]
[a] *University of Eastern Finland, Department of Health and Social Management, Finland*
[b] *Tampere University Hospital, Tampere, Finland*

Abstract. Effective information management promotes safe patient care. Lack of human resources can cause failures when managing patient information. The aim of this study was to analyse the nature of reported patient safety incidents and their location in Choo's information management process model phases when the contributing factor of the incidents was related to human resources and the consequence of the incident was related to harm to the organization's corporate image. Data consisted of the information management related patient safety incident reports (n=475) from 49 health and social care organizations from 2007-2016 in Finland. Deductive analysis and descriptive statistics were used to analyse the data. The results of the study indicated that the shortage of human resources contributed to incomplete documentation, insufficient information sharing between professionals and documenting of information in the wrong place. The majority of the incidents occurred during the information organizing and storage and information distribution phases of the information management process model. Despite the use of electronic health records and electronic patient data, a lack of human resources can lead to breaches in information management processes and harm an organization's corporate image in health and social care contexts.

Keywords. patient safety, information management, human resources

1. Introduction

Health and social care organizations are both information-intensive and information-reliant systems where clinical information collected, recorded, synthesized and shared in different phases of care processes form the core of the information management process [1,2]. In this study, information management is defined according to Choo's information management process model, in which it is formed from a continuous and recurring cycle of six closely related activities: identification of information needs, information acquisition, information organization and storage, development of information products and services, information distribution and information use [3]. In the health and social care context, clinical information needs arise in decision making and care situations [1] requiring the combination of different types of information, such as clinical patient data, personnel expertise, patient's preferences and values and research evidence [4].

[1] Heli Ikonen, Talviunentie 22, 80230 Joensuu, Finland; E-mail: ikonenheli2@gmail.com.

The "Information needs", part of the Choos's information management cycle, is tightly related to information gathering, organizing, and storing [3] through adequate documentation, which can take electronic or paper-based forms such as checklists, care instructions, care plans and handover notes. In Finland, all public hospitals and primary healthcare centres have switched to fully electronic documentation and use of patient data [5]. In the next phase of the information management process, organized and stored information is packaged into different levels of information products [3] in health information systems and electronic health records. Finally, in the phase of "information distribution"[3], written or verbal information is distributed face to face from person to person or through various technologies, systems and methods [6].

Previous studies have found that inadequate information management is a risk for patient safety [7,8]. Safe information management requires that the right information is delivered to the right people in the right place, time and format to be used in care situations [3]. Furthermore, previous studies have indicated that healthcare staffing levels, including staff number or quality, workload and time pressure have effects on patient outcomes such as mortality, medication failures, infections and falls [9, 10, 11]. However, there is a lack of evidence concerning how these factors contribute to patient safety from an information management perspective. The aim of this study was to analyse the nature of reported patient safety incidents and their location in Choo's [3] information management process model phases when the contributing factor of the incidents was related to human resources, and the consequence of the incident was related to harm to a health and social care organization's corporate image.

2. Methods

The data consisted of 110,594 patient safety incident reports between 2007 and 2016 in Finland. In this study all the reports where 1) the reported type on incidence was information management and 2) the reported contributing factor "workload", "staffing number or quality", "shift arrangements and practices" or "time pressure" were included. A total of 475 patient safety incident reports were included. Based on this delineation, research data covered patient safety incident reports from 49 health and social care organizations. Reports were generated via a national, web-based and anonymous patient safety incident reporting system which is used in more than 200 social and health care organizations in Finland. The incident reports included both structured and free-text descriptions of safety events. Deductive analysis based on Choos's information management process model was used for the content of the narrative descriptions of the incident reports (n=475). An analysis matrix with the main categories and variables was created as the basis for classifying and coding the data. Incident report narrative descriptions (n=475) were coded based on the analysis matrix variables, and the coded values were entered into IBM SPSS version 25 (IBM Corporation, USA). Direct distributions and cross tabulation were used to analyse the data.

3. Results

The majority of the reported information management related incidents occurred during the period of active treatment (n=229, 48.2%) or during the interfacility or intrahospital transfer of patients (n=95, 20.0%). Furthermore, 69.3% (n=329) of the incidents were

reported as adverse events, of which 52.4% (n=249) caused no harm to the patient. In most of the incidents, the type of resource shortage was reported as relating to workload, shift arrangements or practices (n=347, 73.1%). As shown in Table 1, most of the 475 reported incidents were classified as mistakes (n=224, 47.2%) or lapses (n=181, 38.1%) as the type of error. In addition, tasks or functions deliberately omitted were represented in 60 (12.6 %) of all the 475 reported incidents. In the research data, failures in decision-making appeared for example as the sending of non-electronic or electronic patient information to the wrong person or place, and as memory-related lapses, such as when the arrangement of patients' follow-on appointments was forgotten by the personnel.

Table 1. Description of the reported patient safety incidents.

Variable	n/475 (%)
Incident location in the treatment process	
Implementation of the treatment	229 (48.2)
Patient transfer	95 (20.0)
Planning or arranging the treatment	86 (18.1)
Follow-up treatment	63 (13.3)
Not known	2 (0.4)
Reported incident type	
Adverse event	329 (69.3)
Near miss	146 (30.7)
Type of error	
Mistake	224 (47.2)
Lapse	181 (38.1)
Undone task	60 (12.6)
Deviation from procedure or failure to perform a task	6 (1.3)
Reported level of consequences of the incidents	
No harm	249 (52.4)
Mild harm	136 (28.6)
Missing	51 (10.7)
Moderate harm	38 (8)
Severe harm	1 (0.2)
Reported type of contributing resource	
Workload, shift arrangements or practices	347 (73.1)
Staff number and quality	128 (26.9)

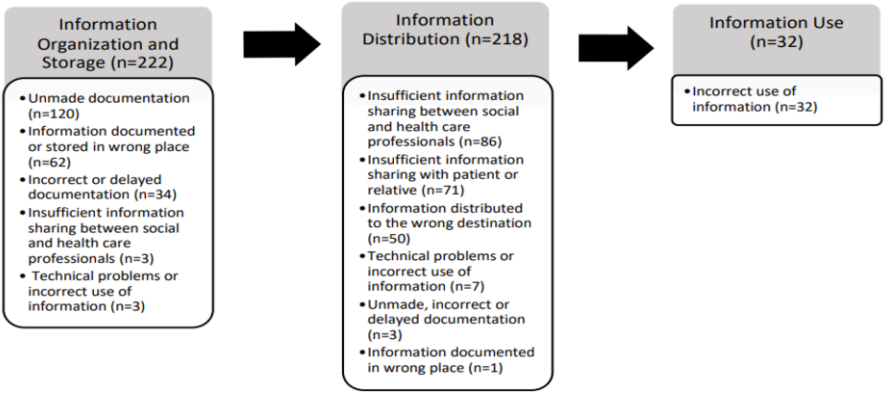


Figure 1. Frequencies of reported patient safety incidents and their position in the modified information management process model [3].

Most of the reported incidents occurred when different types of information were saved or documented (n=222/475, 46.7%), and when verbal or written information was distributed to different locations and to different actors (n=218/475, 45.9%). Hence, the most critical phases in the information management process model [3] for human resource shortage were the information organizing, storage and information distribution phases (Figure 1). Within the information organizing and storage phases, most of the incidents were related to missing documentation (n=120/222) and information documentation or storage in the wrong place (n=62/122). In the information distribution phase, most of the incidents were related to failure in communication between health and social care professionals, or between service users and service providers. Of the 475 reported incidents, 32 took place in the information use phase and were related to incomplete or delayed use or interpretation of information in decision-making and during the patient treatment. A minority (n=3/475, 0.6%) of the reported incidents occurred during the information acquisition and information use phases in the information management process model [3].

4. Discussion

Accurate and intact documenting and distribution of information is a prerequisite for safe and good-quality care processes. This study indicated that a shortage of human resources negatively affects implementation of the treatment, and also patient transfers. Furthermore, the most critical effect of the resource shortage is targeted to failures relating to documentation and communication, occurring particularly as missing documentation, documentation in the wrong place and inadequate or insufficient communication between personnel or between service provider and service user. These findings are supported by the results of previous studies that have highlighted the communication and documentation processes responsible for patient and client safety incidents in various health and social care contexts [6, 7, 8]. The results of this study show similarities with previous research in that from an information management process model perspective [3], information organization and storage and information distribution are the most vulnerable phases for the occurrence of adverse events [7]. The limitations of this study are related to reporting of the patient safety incidents which may be incomplete, and they describe only reporter's perspective of the incident. Further, the criteria on what counts as patient safety incidents may vary at individual and organizational level.

5. Conclusion

Despite the use of electronic health records and electronic patient data, the implementation of information management processes is vulnerable to lack of human resources in health and social care organizations. In particular, documentation and communication in different phases of treatment processes are vulnerable to incidents and therefore the practices related to them need to be critically assessed. Future research should be targeted on other factors, such as IT systems design and usability, contribution to information management related errors in different social and health care settings.

References

[1] Schlegel CS, Yoder L, Jones TL. Clinical Information Needs: A Concept Analysis. Advances in Nursing Science 2020;43(1):E36–45.
[2] MeSH. Health Information Management. Available at: http://www.ncbi.nlm.nih.gov/mesh/68063025. Accessed 22 November 2021.
[3] Choo CW. Information Management for the Intelligent organization. The Art of Scanning the Environment. Third ed. Medford, New Jersey: Information Today, Inc.; 2002. p. 23-42.
[4] AHRQ. Evidence-Based Decisionmaking. Available at: https://www.ahrq.gov/prevention/chronic-care/decision/index.html. Accessed 22 November 2021.
[5] Hyppönen H, Hämäläinen P, Reponen J. E-health and e-welfare in Finland. Check point 2015. Available at: https://www.julkari.fi/handle/10024/129709. Accessed 27 November 2021.
[6] Vermeir P, Vandijck D, Degroote S, Peleman RA, Verhaeghe R, Mortier EP, Hallaert GG, Van Daele S, Buylaert W, Vogelaers D. Communication in healthcare: a narrative review of the literature and practical recommendations. International Journal of Clinical Practice 2015;69(11):1257– 67.
[7] Jylhä V, Bates DW, Saranto K. Critical factors in information management process: the analysis of hospital-based patient safety incident reports. Finnish Journal of eHealth and eWelfare 2016;8(4):164-76.
[8] Kinnunen UM, Kivekäs E, Palojoki S, Saranto K. Register-Based Research of Adverse Events Revealing Incomplete Records Threatening Patient Safety. Digital Personalized Health and Magazine 2020;771 – 5.
[9] Neuraz A, Guérin G, Payet C, Polazzi S, Aubrun F, Dailler F, Lehot J-J, Piriou V, Neidecker J, Rimmelé T, Schott A-M, Duclos A. Patient Mortality is Associated with Staff Resources and Workload. A Multicenter Observational Study. Critical Care Medicine 2015;43(8):1587-94.
[10] Cho E, Chin DL, Kim S, Hong O. The Relationship of Nurse Staffing Level and Work Environment with Patient Adverse Events. Journal of nursing scholarship 2016;48(1):74-82.
[11] Driscoll A, Grant MJ, Carroll D, Dalton S, Deaton C, Jones I, Lehwaldt D, McKee G, Munyombwe T, Astin F. The effect of nurse-to-patient ratios on nurse-sensitive patient outcomes in acute specialist units: a systematic review and meta-analysis. European journal of cardiovascular nursing 2018;17(1):6-22.

Challenges of Trustable AI and Added-Value on Health
B. Séroussi et al. (Eds.)

doi:10.3233/SHTI220430

Making EHRs Trustable: A Quality Analysis of EHR-Derived Datasets for COVID-19 Research

Miguel PEDRERA-JIMENEZ[a,b,1], Noelia GARCIA-BARRIO[a], Paula RUBIO-MAYO[a], Guillermo MAESTRO[c], Antonio LALUEZA[c], Ana GARCIA-REYNE[c], María José ZAMORRO[c], Alejandra PONS[c], María Jesús SANCHEZ-MARTIN[c], Jaime CRUZ-ROJO[a], Víctor QUIROS[a], José María AGUADO[c], Juan Luis CRUZ-BERMUDEZ[a], José Luis BERNAL[a], Laura MERSON[d], Carlos LUMBRERAS[c] and Pablo SERRANO[a]
[a]*Data Science Unit, Research Institute Hospital 12 de Octubre, Madrid, Spain.*
[b]*ETSI Telecomunicación, Universidad Politécnica de Madrid, Madrid, Spain.*
[c]*Internal Medicine Department, Hospital Universitario 12 de Octubre, Madrid, Spain.*
[d]*ISARIC Global Support Centre, University of Oxford, Oxford, UK.*

Abstract. One approach to verifying the quality of research data obtained from EHRs is auditing how complete and correct the data are in comparison with those collected by manual and controlled methods. This study analyzed data quality of an EHR-derived dataset for COVID-19 research, obtained during the pandemic at Hospital Universitario 12 de Octubre. Data were extracted from EHRs and a manually collected research database, and then transformed into the ISARIC-WHO COVID-19 CRF model. Subsequently, a data analysis was performed, comparing both sources through this convergence model. More concepts and records were obtained from EHRs, and PPV (95% CI) was above 85% in most sections. In future studies, a more detailed analysis of data quality will be carried out.

Keywords. Electronic Health Records, Real World Data, Data Quality, Completeness, Correctness, Semantics, Standards, ISARIC-WHO, COVID-19.

1. Introduction

Electronic health records (EHRs) are conceived as a digital repository of health data that is used for individual patient healthcare [1]. In addition, it can be applied to other purposes, known as secondary uses, including clinical research and public health [2]. To achieve this, recent studies have designed methodologies based on health information standards for allowing the effective reuse of EHRs, which is essential for obtaining data in an agile, flexible and efficient way [3, 4]. Traditionally, data for research have been manually collected in purpose-built and controlled databases. This made it is necessary to audit the quality of data obtained from EHRs through systematic and validated methodologies [5, 6]. This became evident during the COVID-19 pandemic when two studies, published in high-impact journals, had to be retracted due to data quality, among other issues (10.1016/S0140-6736(20)31180-6 and 10.1056/NEJMoa2007621).

[1] Corresponding Author, Miguel Pedrera Jiménez, Data Science Group, Hospital Universitario 12 de Octubre, Av. de Andalucía S/N, 28041, Madrid, España; E-mail: miguel.pedrera@salud.madrid.org.

Thus, this study analyzes data quality of an EHR-derived dataset for COVID-19 research, obtained at Hospital Universitario 12 de Octubre, Madrid, Spain (H12O) [7].

2. Methods

The study compares two study databases: one obtained only from structured EHRs, and other manually transcribed from structured EHRs, clinical reports, and external sources. Data for both databases were collected between March 2020 and September 2020. This work is part of the research line on reuse of EHRs at H12O [3, 4, 7-9].

2.1. Extraction, transformation and loading of health data

The first source for data extraction was the structured EHRs of H12O, which have been modeled and formalized through health information standards including ISO 13606 [10], and controlled terminologies such as SNOMED CT and LOINC [11, 12]. This effort has allowed the full meaning reuse of EHRs, without additional manual efforts, in data collection processes for research [3]. Hence, it was possible to participate in different data-driven projects during the COVID-19 pandemic, including TriNetX, EHDEN Consortium, 4CE Consortium and ISARIC Consortium [4, 7].

The second data source was the data collected within the STOP-CORONAVIRUS project, a clinical study on COVID-19 developed at H12O [13]. In this study, a specific data collection was carried out in a relational database implemented in REDCap [14], being manually transcribed from structured EHRs, clinical reports and external sources.

Data from both sources were transformed into a common model for analysis and comparison. The COVID-19 case report form (CRF) proposed by ISARIC-WHO was chosen as convergence model, due to its international adoption for COVID-19 clinical data collection, and because H12O participates in the ISARIC Consortium by transferring EHRs without manual efforts. [15, 16]. Thus, two identical databases based on this model were implemented in REDCap [14], and then the data from the same cohort of 1732 COVID-19 patients were loaded into them from each data source.

2.2. Data quality analysis of health data

Determining the quality of health data is not a straightforward task since it can be measured from different perspectives. In the review carried out by Weiskopf et al., five data quality dimensions were identified, as well as seven methods to evaluate them [5]. Based on this review, the 'gold-standard comparison' method was selected for the analysis, establishing the STOP-CORONAVIRUS dataset as the reference against which to compare data obtained from EHRs. Two data quality dimensions were analyzed:

- **Completeness**, i.e., if a fact about a patient is recorded in EHRs. For this, both data sources were analyzed to determine the coverage of the concepts specified by ISARIC-WHO, the volume of data obtained (patient-level records), and the cohort coverage achieved.
- **Correctness**, i.e., if a record present in EHRs is true. For this, both data sources were compared to determine, for each equivalent (patient, date of registration and concept) non-null record, whether EHRs report same content as the gold-standard. Thus, the Positive Predictive Value (PPV, 95% CI) was calculated.

Hence, an algorithm was implemented with R programming language [17], with which it was determined, for each patient and record, if the data exists in both sources, and if so, whether they match (same data type and content). Figure 1 shows the flowchart of the algorithm.

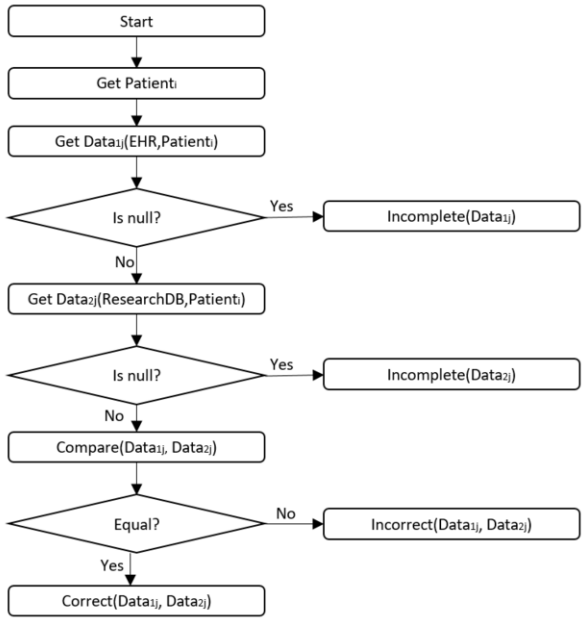


Figure 1. Algorithm for data quality analysis for completeness and correctness dimensions.

3. Results

3.1. Completeness analysis

The first result obtained was the completeness analysis of EHRs vs. the research database. Table 1 shows, grouped by section of the ISARIC-WHO CRF, the percentage of the CRF concepts and cohort covered, and the absolute volume of records obtained.

Table 1. Completeness analysis results.

ISARIC CRF Section	EHRs			Research database		
	Concepts (%)	Cohort (%)	Records (N)	Concepts (%)	Cohort (%)	Records (N)
Demographics	42.11	99.88	10,877	26.32	100	8051
Onset & admission	100	99.88	5266	82.35	100	3420
Signs and symptoms	53.85	99.88	12,714	30.77	100	20,177
Pre-admission medication	58.33	99.88	12,110	33.33	100	5318
Co-morbidities	100	99.88	34,600	75	100	24,948
Treatment	80	99.88	17,886	64	100	15,307
Complications	100	99.88	50,170	88.89	100	13,727
Clinical diagnostics	53.33	99.88	10,474	53.33	100	9998
Microbiology diagnostics	66.67	99.88	13,644	66.67	92.44	8511
Medication	82.35	99.88	27,554	61.76	100	20,285
Daily observations	86.96	99.88	331,796	30.43	99.48	27,848
Outcome	100	99.88	5190	66.67	100	3377

3.2. Correctness analysis

The second result obtained was the analysis of the correctness of EHRs compared with those recorded manually in the research database. Table 2 shows, for each section of the ISARIC-WHO CRF, the PPV (95% CI), as well as the number of concepts and records that could be compared.

Table 2. Correctness analysis results.

ISARIC CRF Section	Concepts compared	Records compared	PPV (95% CI)
Demographics	5	7505	97.79 (97.46, 98.12)
Onset & admission	2	1793	91.02 (89.70, 92.34)
Signs and symptoms	9	5824	58.41 (57.14, 59.68)
Pre-admission medication	4	5316	91.99 (91.26, 92.72)
Co-morbidities	15	24,933	91.36 (91.01, 91.71)
Treatment	15	13,805	84.80 (84.20, 85.40)
Complications	8	13,715	88.05 (87.51, 88.59)
Clinical diagnostics	5	5384	84.77 (83.81, 85.73)
Microbiology diagnostics	4	5008	99.16 (98.91, 99.41)
Medication	19	19,336	73.29 (72.67, 73.91)
Daily observations	14	19,690	78.70 (78.13 ,79.27)
Outcome	2	3375	92.74 (91.86, 93.62)

4. Discussion

The completeness analysis showed that, although the cohort was mostly covered from both sources, more concepts and volume of records were obtained from EHRs, since they were recorded during and for patient's care, rather than in an additional effort based on a fixed design. This was most evident in the 'Daily observations' section, with 86.96% concepts and 331,796 records in EHRs, vs. 30.43% and 27,848 in the research database.

Likewise, the correctness analysis showed that the EHRs has a PPV (95% CI) over 85% in most sections. The sections 'Signs and Symptoms', 'Medication' and 'Daily observations' were between 58% and 80%, which, thanks to this analysis, could be identified as gaps in the standardization and coverage of the EHRs (mainly due to free-text data entry). This allowed improving these information domains and thus obtaining higher quality data in future processes of obtaining EHRs-derived datasets for research.

These results highlight the value of the EHRs as a useful and valid source for research, in a scenario where multiple projects propose automated upload processes from healthcare information systems to databases and repositories for research [18, 19].

5. Conclusions

In this study, a quality analysis of EHR-derived datasets for COVID-19 research was performed. To this end, data were extracted from two sources: EHRs of H12O and a manually-collected research database. Then, both datasets were transformed into the ISARIC-WHO COVID-19 CRF model for comparative analysis. Thus, it could be concluded that the EHRs are more complete than a specific research database, and these data collected during the healthcare activity have an adequate accuracy.

In future studies, expanded and more detailed analysis will be performed, including the results for each of the concepts of the ISARIC-WHO CRF model.

Acknowledgment

This work has been supported by Research Projects PI18/00981, PI18/01047 and COV20/00181; funded by Instituto de Salud Carlos III, co-funded by ERDF/ESF.

We want to thank the ISARIC Consortium for its innovative vision in accepting our proposal for automated data upload from EHRs. We also want to thank Mª Teresa García Morales and Agustín Gómez de la Cámara (Scientific Support Unit, "i+12" Institute).

References

[1] Häyrinen K, Saranto K, Nykänen P. Definition, structure, content, use and impacts of electronic health records: a review of the research literature. *Int J Med Inform*. 2008;77(5):291-304. doi:10.1016/j.ijmedinf.2007.09.001.

[2] Safran C, Bloomrosen M, Hammond WE, et al. Toward a national framework for the secondary use of health data: an American Medical Informatics Association White Paper. *J Am Med Inform Assoc*. 2007;14(1):1-9. doi:10.1197/jamia.M2273

[3] Pedrera-Jiménez M, García-Barrio N, Cruz-Rojo J, et al.Obtaining EHR-derived datasets for COVID-19 research within a short time: a flexible methodology based on Detailed Clinical Models. *J Biomed Inform*. 2021;115:103697. doi:10.1016/j.jbi.2021.103697.

[4] Pedrera M, Garcia N, Rubio P, Cruz JL, Bernal JL, Serrano P. Making EHRs Reusable: A Common Framework of Data Operations. *Stud Health Technol Inform*. 2021;287:129-133. doi:10.3233/SHTI210831.

[5] Weiskopf NG, Weng C. Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research. *J Am Med Inform Assoc*. 2013;20(1):144-151. doi:10.1136/amiajnl-2011-000681.

[6] Kohane IS, Aronow BJ, Avillach P, et al. What Every Reader Should Know About Studies Using Electronic Health Record Data but May Be Afraid to Ask. *J Med Internet Res*. 2021;23(3):e22219. Published 2021 Mar 2. doi:10.2196/22219.

[7] Pedrera M, Garcia N, Blanco A, et al. Use of EHRs in a Tertiary Hospital During COVID-19 Pandemic: A Multi-Purpose Approach Based on Standards. *Stud Health Technol Inform*. 2021;281:28-32. doi:10.3233/SHTI210114.

[8] Pedrera M, Serrano P, Terriza A, et al. Defining a Standardized Information Model for Multi-Source Representation of Breast Cancer Data. *Stud Health Technol Inform*. 2020;270:1243-1244. doi:10.3233/SHTI200383.

[9] González L, Pérez-Rey D, Alonso E, et al. Building an I2B2-Based Population Repository for Clinical Research. *Stud Health Technol Inform*. 2020;270:78-82. doi:10.3233/SHTI200126.

[10] ISO 13606 standard. https://www.iso.org/standard/67868.html. Accessed January 7, 2022.

[11] Donnelly K. SNOMED-CT: The advanced terminology and coding system for eHealth. *Stud Health Technol Inform*. 2006;121:279-290.

[12] McDonald CJ, Huff SM, Suico JG, et al. LOINC, a universal standard for identifying laboratory observations: a 5-year update. *Clin Chem*. 2003;49(4):624-633. doi:10.1373/49.4.624

[13] STOP-CORONAVIRUS. https://imas12.es/blog/stop-coronavirus-nuevo-proyecto-clinico-llevado-a-cabo-en-el-instituto-i12-para-ofrecer-respuestas-integrales-a-la-covid-19/. Accessed January 7, 2022.

[14] Harris PA, Taylor R, Minor BL, et al. The REDCap consortium: Building an international community of software platform partners. *J Biomed Inform*. 2019;95:103208. doi:10.1016/j.jbi.2019.103208R.

[15] ISARIC-WHO CRF for COVID-19. https://isaric.org/research/covid-19-clinical-research-resources/covid-19-crf/. Accessed January 7, 2022.

[16] ISARIC Clinical Characterisation Group. The value of open-source clinical science in pandemic response: lessons from ISARIC [published correction appears in Lancet Infect Dis. 2021 Dec;21(12):e363]. *Lancet Infect Dis*. 2021;21(12):1623-1624. doi:10.1016/S1473-3099(21)00565-X.

[17] Foundation for Statistical Computing. https://www.r-project.org/about.html. Accessed January 7, 2022.

[18] Maldonado JA, Marcos M, Fernández-Breis JT, et al. CLIN-IK-LINKS: A platform for the design and execution of clinical data transformation and reasoning workflows. *Comput Methods Programs Biomed*. 2020;197:105616. doi:10.1016/j.cmpb.2020.105616.

[19] Ong TC, Kahn MG, Kwan BM, et al. Dynamic-ETL: a hybrid approach for health data extraction, transformation and loading. *BMC Med Inform Decis Mak*. 2017;17(1):134. Published 2017 Sep 13. doi:10.1186/s12911-017-0532-3.

Challenges of Trustable AI and Added-Value on Health
B. Séroussi et al. (Eds.)

doi:10.3233/SHTI220431

Usability Assessment of Conversational Agents in Healthcare: A Literature Review

Kerstin DENECKE[a,1] and Richard MAY[b]
[a] *Bern University of Applied Sciences, Bern, Switzerland*
[b] *Harz University of Applied Sciences, Wernigerode, Germany*

Abstract. Conversational agents (CA) are chatbot-based systems supporting the interaction with users through text, speech, or other modalities. They are used in an increasing number of medical use cases. Even though usability is considered a prerequisite for the success of mHealth apps using CA, there is still no standard procedure to study usability of health CA. In this paper, we report the results from a systematic literature review aiming at identifying study designs, tools, and metrics used to assess usability in health CA. We searched three bibliographic databases (PubMed, Scopus, IEEE Xplore) for papers reporting on CA in healthcare to extract information on the usability assessment of those CA. From 273 retrieved results, we included 66 papers for full text review. 34 of them reported on usability assessments. A broad range of tools is used (e.g. SUS, UEQ), but also individual questionnaires are exploited. The examined studies use scenario-based setups but assess also real-world usage. Exploratory setups are rarely reported. Due to the differences in the study designs and assessment tools, it is impossible to compare usability among CA. Thus, we recommend to develop a standardised procedure that can be always applied and which can be enriched by assessments needed for evaluating usability of CA-specific features.

Keywords. Conversational agents, usability, evaluation, chatbot, healthcare

1. Introduction

With the advent of artificial intelligence and its use for understanding and interpreting speech, conversational agents (CA) and their application in healthcare have gained enormous interest in recent years. CA are chatbot-based systems supporting the interaction with users through text, speech, or other modalities in a variety of medical use cases, such as triage systems [1], for medication management [2], or to recommend ICD-10 codes [3]. CA allow patients to receive immediate response (e.g. when having questions on the medication) or facilitate human-machine interaction. Equipped with empathy features, CA can create a bond of trust with its user which is impossible for other IT systems.

Usability is considered a prerequisite for the success of any mHealth app as it is one of the main indicators for the overall acceptance and success of an application. Moreover, it is a quality attribute that assesses how easy user interfaces are to use [4]. It is typically measured by having a number of test users using the system to perform a prespecified set of tasks. However, despite the extensive research on CA in healthcare, we are not

[1] Corresponding Author, Kerstin Denecke, Institute for Medical Informatics, Bern University of Applied Sciences, Quellgasse 21, 2501 Biel, Switzerland; E-mail: kerstin.denecke@bfh.ch

aware of a systematic overview of CA-related usability assessment. In this paper, we aim to assess recent research on usability assessment in the context of CA, focusing on tools and evaluation metrics used. We will discuss best practices and open research avenues. Finally, a replication package containing lists of all selected publications, including our extraction sheet will be delivered.

2. Methodology

To achieve our research objectives, we employed a systematic literature review according to the guidelines of Kitchenham et al. [5]. In the following subsections, we describe the individual steps of our study based on these guidelines.

Initial Screening. First, we conducted an initial screening to ensure that there is a relevant body-of-knowledge in the field of CA in healthcare, i.e. there is sufficient research interest. For this purpose, we applied the following search string to the literature databases Pubmed, IEEE Xplore, and Scopus: ("chatbot" OR "conversational agent") AND ("healthcare" OR "health care"). This search string resulted in a total of 720 results (103 on IEEE Xplore, 143 on Pubmed, 474 on Scopus). Thus, we confirmed considerable research interest in CA in healthcare and considered our research objective valuable enough to initiate our detailed systematic literature study.

Search String. Second, we performed the refined literature search on Pubmed, IEEE Xplore, and Scopus based on the general findings of our initial screening. These databases include peer-reviewed literature of diverse publishers, which reduced the threat of missing relevant papers and ensured a high publication quality. In this context, we applied the following search string: *("conversational user interface" OR "intelligent agent" OR "conversational agent" OR chatbot) AND (health OR healthcare OR medicine) AND evaluation.* In the Pubmed search string, we resisted on including search terms "(health OR healthcare OR medicine)" since Pubmed basically contains literature from the health domain.

Selection Criteria (SC): To identify relevant papers, we used the following criteria:

- **SC1**: The publication is written in English.
- **SC2**: The publication is a peer-reviewed conference paper or journal article.
- **SC3**: The publication has been published between 2012 and 2021.
- **SC4**: The publication is longer than five pages.
- **SC5**: The publication reports on studies dealing with CA.

We intentionally focused on the last decade to cover the most recent research (**SC3**). We argue that relevant findings on CA older than ten years have typically already become established fundamentals or practices. **SC4** was used to ensure a certain quality of the papers, assuming that a publication with a minimum number of pages provides enough details to comprehend the addressed problem. Moreover, we relied on the review of publication venues by the chosen literature databases. This is a well-established adaptation, as we structure previous findings based on different research methods [7]. Regarding **SC5**, review papers as well as papers reporting theoretical frameworks or complete conference proceedings were excluded. Further, we removed publications dealing with embodied CA or systems that send only push up notifications as they are based on other requirements than traditional text- or voice-based CA. Systems not dealing with healthcare were also removed.

Data Extraction: To extract relevant data, we defined the following criteria: type of CA, e.g. coaching or informational, input-output type, e.g. text or speech, evaluation

aspect, e.g. usability or efficacy, number of participants, type of participants, e.g. students or patients, metrics used, e.g. SUS or UEQ, tools used, e.g. questionnaire or interview, and evaluation execution, e.g. scenario or exploratory.

Conduct: We conducted the literature search on December 12, 2021. Overall, we identified 273 results (24 on IEEE Xplore, 36 on Pubmed, 213 on Scopus). In a first step, both authors manually reviewed the papers by using the collaborative review tool Rayyan QCRI which automatically removed 16 duplicates. Each reviewer examined half of the publications' titles and abstracts resulting in the exclusion of 186 papers including nine additional duplicates. Next, the full texts were downloaded for detailed analyses. However, for three papers the full texts were not accessible, i.e. we considered 84 papers in the full text review. All disagreements between the authors were resolved during discussions until a consensus on a decision was achieved.

3. Results

Data assessment***:*** In the assessment phase, 18 papers were removed because they did not fulfil the inclusion criteria (embodied CA, no evaluation results described or CA in domains other than health). 66 papers were finally assessed, and data extracted. The complete reference list and the data extraction sheet is available online[2]. The CA included in our analysis addressed multiple application areas: disease management (asthma, diabetes, chronic pain), intelligent interviewer (family history, PROM), retrieval (for physicians ICD-10 encoding, for patients' information retrieval), mental health (mainly delivering cognitive behavioural therapy for different mental disorders or patient education), medication management.

CA and its conversation-related characteristics: Based on our previous work [6], we assume four different types of CA: informational, coaching, questioning, and monitoring. In the analysed papers, most CA focused on coaching tasks (62.1%). 19.7% of the papers described questioning CA. We also identified that 18.2% of the CA only provide information. 4.6% of the studies presented monitoring CA. Overall, we found out that three CA (4.6%) perform a higher number of different tasks than the other CA, making it impossible to assign them to a single CA type. However, we could not find any relationships between these CA. Regarding the input/output type, all CA are at least based on text. Even 78.8% of the CA are solely based on text. Furthermore, 18.2% of the studies presented CA based on a combination of text and voice. In one paper, a voice user interface without text was described. Another CA combined specific visual elements with text.

Evaluation aspects. About half of the studies considered usability in their CA evaluation (51.5%). A similar number of papers also evaluated user experience, including user satisfaction or human-machine interaction (47.0%). We also identified that 39.4% of the CA were evaluated on technical aspects, including effectiveness, efficiency, performance, or reliability. 16.7% of the papers evaluated the (subjective) effectiveness of the respective CA-related intervention actions on the user. In addition, 9.1% of the studies indicated the technological acceptance of the CA. Overall, 21.2% of the papers reported only one evaluation criterion, such as user experience. However, we assume that these are usually generic terms for several partial evaluation aspects.

[2] https://github.com/Rim007/ReplicationPackage_UsabilityCA

Characteristics of the usability tests: Since our research focuses on the usability evaluation of CA, in the following we will only discuss those papers that actually used usability as an evaluation criterion (n=34). We found out that 11.8% of the usability studies were conducted with 1 to 10 participants. Almost half of all evaluations were performed with 11 to 50 participants (47.0%). Furthermore, 17.7% of the papers presented evaluations with 51 to 100 and 20.6% with more than 100 participants. One paper did not report a number. We could distinguish 5 different groups of participants: doctors / therapists / experts were involved in testing (11.8%), students / staff members (20.6%), random users (20.6%), patients (41.2%) and children/teenagers (5.9%). One paper did not report about the participants (2.9%). One paper involved patients and physicians. 91.1% of the studies exploited questionnaires to assess the usability; 14.7% of the papers reported on interviews and 11.8% analysed the protocols of the conversation to learn about user behaviour and usability. Some of the studies exploited existing usability questionnaires; among them are User experience questionnaire [7-9], (n=3), System Usability Scale [10-12] (n=4), Net promoter score (n=1), Subjective Assessment of System Speech Interfaces [13] (n=1), FEDS [14] (n=1), User engagement scale [11] (n=4), TRINDI [2] (n=1), PARADISE framework [2] (n=1), ISO 9214 [15,16] (n=2), UTAUT [17] (n=1). When not relying upon existing questionnaires, the researcher developed own surveys, often comprising only few questions (e.g. "is it easy to use?"). The main usability test setup was a real-world application, i.e. the system was used in daily practice by participants (38.2%), or a scenario-based usability test (44.1%). Users were asked to explore the application on their own in 2 studies (5.9%). 4 studies (11.8%) did not provided any information on the setup.

4. Discussion and conclusion

Based on the results, we found that most of the examined CA are based on text input/output aiming at coaching their users in a specific domain. On average, around 55 participants were involved in a scenario-based usability evaluation, which indicates quantitative assessment. When evaluating CA, user experience is usually examined besides usability. This shows the strong link between both aspects. However, usability is usually associated with more technical aspects than user experience, which addresses more subjective aspects, such as user satisfaction. Therefore, we assume that there is an insufficient understanding of evaluation criteria, such as the differentiation between usability and user experience. This issue is also supported by the fact that generic terms, such as interaction, were sometimes used as the only, subjective evaluation criteria. This indicates that some studies are less concerned with the evaluation of certain aspects than with the evaluation itself. However, this considerably impairs the comprehensibility of the evaluation method and its results. Only one paper [13] used an assessment tool specifically designed for assessing speech interfaces. The other tools have originally been developed for IT systems in general, not specifically for CA. The comparison of Holmes et al. showed that conventional tools like SUS are not as accurate when applied to CA [18]. We strongly recommend harmonising usability evaluation strategies to create a uniform understanding and basis for usability assessment and thus enable a comparison of evaluation results. For example, an agreed scoring system specifically developed for CA in healthcare to be used in all usability testings would allow comparison. In healthcare, we have to deal with diverse user groups covering multiple social dimensions and diverse levels of cognitive capabilities in a variety of use cases. These specific

requirements cannot be completely addressed by the established assessment tools. Further, the results showed that diversity still remains unconsidered in usability testing. Since communication is language-based with CA, assessing usability in correlation with user's language and communication skills is relevant. To address these limitations, we plan to develop a suitable evaluation tool to provide a uniform understanding and an essential basis for the comprehensible usability evaluation of healthcare-related CA. However, to create this tool, we have to answer the following research questions in future research: Would an exploratory design of a usability test be more appropriate? Would interviews with participants reveal additional challenges? What can we learn from the conversation flow on usability?

References

[1] Omoregbe NAI, et al.. Text Messaging-Based Medical Diagnosis Using Natural Language Processing and Fuzzy Logic. Dogra A, editor. J Health Eng. 2020 Sep 29;2020:1–14.

[2] Hess GI, Fricker G, Denecke K. Improving and Evaluating eMMA's Communication Skills: A Chatbot for Managing Medication. Stud Health Technol Inform. 2019;259:101–4.

[3] Siangchin N, Samanchuen T. Chatbot Implementation for ICD-10 Recommendation System. In: 2019 Intern Conf on Engineering, Science, and Industrial Applications (ICESI). Tokyo, Japan; 2019. p. 1–6.

[4] Wilson C, editor. User experience re-mastered: your guide to getting the right design. Burlington, MA: Morgan Kaufmann Publishers; 2010. 382 p.

[5] Kitchenham BA, Budgen D, Brereton P. Evidence-based software engineering and systematic reviews. Boca Raton: CRC Press; 2016. 399 p.

[6] May R, Denecke K. Security, privacy, and healthcare-related conversational agents: a scoping review. Inform Health Soc Care. 2021 Oct 7;1–17.

[7] Denecke K, Vaaheesan S, Arulnathan A. A Mental Health Chatbot for Regulating Emotions (SERMO) - Concept and Usability Test. IEEE Trans Emerg Top Comput. 2020;1–1.

[8] Chatzimina M, Koumakis L, Marias K, Tsiknakis M. Employing Conversational Agents in Palliative Care: A Feasibility Study and Preliminary Assessment. In: 2019 IEEE 19th International Conference on Bioinformatics and Bioengineering (BIBE). Athens, Greece: IEEE; 2019. p. 489–96.

[9] Chatzimina M, et al. Designing a conversational agent for patients with hematologic malignancies: Usability and Usefulness Study. In: 2021 IEEE EMBS International Conference on Biomedical and Health Informatics (BHI). Athens, Greece: IEEE; 2021. p. 1–4.

[10] Kadariya D, Venkataramanan R, Yip HY, Kalra M, Thirunarayanan K, Sheth A. kBot: Knowledge-Enabled Personalized Chatbot for Asthma Self-Management. In: 2019 IEEE International Conference on Smart Computing (SMARTCOMP). Washington, DC, USA: IEEE; 2019. p. 138–43.

[11] Oh J, Jang S, Kim H, Kim J-J. Efficacy of mobile app-based interactive cognitive behavioral therapy using a chatbot for panic disorder. Int J Med Inf. 2020 Aug;140:104171.

[12] Bennion MR, Hardy GE, Moore RK, Kellett S, Millings A. Usability, Acceptability, and Effectiveness of Web-Based Conversational Agents to Facilitate Problem Solving in Older Adults: Controlled Study. J Med Internet Res. 2020 May 27;22(5):e16794.

[13] Yasavur U, Lisetti C, Rishe N. Let's talk! speaking virtual counselor offers you a brief intervention. J Multimodal User Interfaces. 2014 Oct;8(4):381–98.

[14] Meier P, Beinke JH, Fitte C. FeelFit - Design and Evaluation of a Conversational Agent to Enhance Health Awareness. Int Conf Inf Syst 2019 ICIS 2019. 2019;

[15] Ghaleb M, Almurtadha Y, Algarni F, Abdullah M, Felemban E, M. Alsharafi A, et al. Mining the Chatbot Brain to Improve COVID-19 Bot Response Accuracy. Comput Mater Contin. 2022;70(2):2619–38.

[16] Sia DE, Yu MJ, Daliva JL, Montenegro J, Ong E. Investigating the Acceptability and Perceived Effectiveness of a Chatbot in Helping Students Assess their Well-being. In: Asian CHI Symposium 2021. Yokohama Japan: ACM; 2021. p. 34–40.

[17] Valtolina S, Hu L. Charlie: A chatbot to improve the elderly quality of life and to make them more active to fight their sense of loneliness. In: CHItaly 2021: 14th Biannual Conference of the Italian SIGCHI Chapter. Bolzano Italy: ACM; 2021. p. 1–5.

[18] Holmes S, Moorhead A, Bond R, Zheng H, Coates V, Mctear M. Usability testing of a healthcare chatbot: Can we use conventional methods to assess conversational user interfaces? In: Proceedings of the 31st European Conference on Cognitive Ergonomics. BELFAST United Kingdom: ACM; 2019. p. 207–14.

Challenges of Trustable AI and Added-Value on Health
B. Séroussi et al. (Eds.)

doi:10.3233/SHTI220432

Metadata Definition in Registries: What Is a Data Element?

Jürgen STAUSBERG[a,1], Sonja HARKENER[a], Markus BURGMER[b], Christoph ENGEL[c], Robert FINGER[d], Carsten HEINZ[e], Ekkehart JENETZKY[f,g], David MARTIN[f,h], Rüdiger RUPP[i], Martin SCHOENTHALER[j], Christian SCHULD[i], Barbara SUWELACK[k] and Jeannine WEGNER[k]

[a] *University Duisburg-Essen, Faculty of Medicine, IMIBE, Essen, Germany*
[b] *Department of Psychosomatics and Psychotherapy, LWL-Hospital and University Hospital Münster, Münster, Germany*
[c] *Leipzig University, IMISE, Leipzig, Germany*
[d] *Department of Ophthalmology, University Hospital Bonn, Bonn, Germany*
[e] *Department of Ophthalmology, St. Franziskus-Hospital Münster, Münster, Germany*
[f] *Faculty of Health/School of Medicine, Witten/Herdecke University, Witten, Germany*
[g] *Department of Child and Adolescent Psychiatry and Psychotherapy, University Medical Center of the Johannes-Gutenberg-University, Mainz, Germany*
[h] *Department of Pediatrics, Eberhard-Karls University Tübingen, Tübingen, Germany*
[i] *Spinal Cord Injury Center, Heidelberg University Hospital, Heidelberg, Germany*
[j] *Department of Urology, University of Freiburg, Freiburg, Germany*
[k] *University Hospital Münster, Transplant Nephrology, Münster, Germany*

Abstract. Observational research benefits from a rich methodological foundation of registry development and operation published in international and national guidelines. Metadata management is an essential part of registry implementation based on concepts of data elements and value sets. The metadata from six German registries revealed vastly divergent interpretations of the concept of data elements. The different perspectives of research questions, data acquisition and data storage were all represented in the registries' catalogs of data elements. Consequently, the whole life cycle of a registry needs to be accompanied by a catalog of data elements, which has to be continuously adapted to the changing perspectives. A standard for the representation of those metadata is still missing. The FAIR Guiding Principles introduce important methodological requirements, but the tools for their fulfillment in respect to the management of metadata are still in its infancy.

Keywords. Data element, documentation, metadata, registries.

1. Introduction

Registries are an established method in observational studies in health services research, in epidemiology and public health, and in clinical research. Dating back to the National Leprosy Registry of Norway [1] established in 1856, registry research is nowadays in a

[1] Corresponding Author, Jürgen STAUSBERG, Institute for Medical Informatics, Biometry and Epidemiology, Faculty of Medicine, University Duisburg-Essen, Hufelandstrasse 55, 45122 Essen, Germany; E-mail: stausberg@ekmed.de.

consolidated phase with a rich set of standards and recommendations for best practices [2]. The management of metadata, i.e. mainly the definition of data elements and value sets, represents an essential task within development and operation of a registry. However, it is a real challenge to reduce strongly or to remove the mess of concepts and terms related to metadata [3].

There is consent in international and national guidelines about research questions as starting points for the definition of the registries' metadata. The US-American Agency for Healthcare Research and Quality (AHRQ) describes the translation of these questions into measurable exposures and outcomes as an essential step of implementation [4]. These measures are represented by data elements that are grouped to subcategories and assigned to one of the three domains characteristics, treatments, and outcomes. The AHRQ further points out that data elements can also be defined from the perspective of data analyses, e.g. by calculating new data elements from others as the body mass index from body weight and height. Furthermore, a data element can be associated to a source, for example a case report form (CRF), documenting the type of the source or the time of recording [5]. Unfortunately, a standardized organization structure of registries' metadata with the levels measures, analysis and source is neither explicitly consented in registry research [6] nor reflected by the third edition of the international standard ISO/IEC 11179-3 "Information technology - Metadata registries (MDR)" [7].

A consensus approach to the definition of registry data elements was the aim of this work, because of the lack of consensus in this definition.

2. Methods

2.1. Material

Six registries were included in a survey within a funding initiative for health services research (cf. [8] for details). The registries covered different fields of health care, lifelong monitoring of people with spinal cord injury (A), treatment exit options for uveitis (B), hereditary breast and ovarian cancer (C), safety of living kidney donors (D), recurrent urolithiasis of the upper urinary tract (E), and fever in childhood (F). In order to perform a cross-project comparison, a template based on Microsoft Access for a common representation of the individual metadata was forwarded to the registries [9].

Three projects reported the metadata of their registries based on this template (A, B, C). One delivered a codebook exported from an electronic data capture system (D); another provided a set of text-files based on Microsoft Word with metadata organized in tables (E). The sixth project used a proprietary tabular format (F). Only one project delivered their metadata fully compliant with the template's format (A). Consequently, 1) all material was transferred by the independent group mentioned above into the template's structure without changing the project's interpretation of the concept "data element" and 2) information was added where necessary, e.g. a documentation object as an umbrella structure for data elements. The complete list of metadata was finally embedded into the meta model of ISO/IEC 11179-3 for further analyses [10].

2.2. Reference system

Figure 1 shows the reference system applied in this work. For answering research questions, information needs to be collected to allow analyses. The information is

formally represented in a catalog of data elements [11], independently from technical aspects related to data storage or data acquisition. The catalog of data elements defines the structure of the data storage as well as the content of CRFs or other means of data acquisition. Reusing data from existing sources will possibly set constraints for the latter.

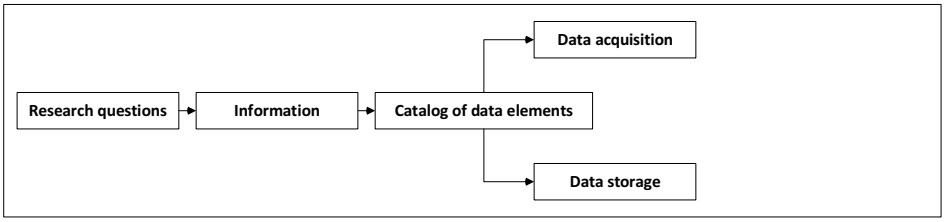


Figure 1. Reference system for metadata of registries.

3. Results

Registry A delivered the catalog of data elements as requested. Data elements represented information independently from technical constraints through the storage or the acquisition of data. Data elements were designated with generally understandable terms, e.g. “gender” or “date of birth”. In contrary, registry C filled the designations of data elements with shortcuts like “SEX” or “BIRDAT”, called labels in the following. The relationship from labels to information was then provided through a description in the template, however with a non-uniform format. Registry B, also delivering the metadata with the template, partially used questions as data elements, like “What was the date of birth?” Furthermore, registry B added context to a data element in its denomination, for example the role of the person that is asked for his or her date of birth.

The codebook from registry D contained names of variables and fields. The entries were mainly short forms of designations like “dateofbirth”, complemented through a counter as “drug_1” where needed. The counters were related to multiple answers as well as to multiple time points. Registry E used the full capacity of free text, e.g. by defining more than one data element in a table cell. Such data element collections will be divided for data storage and data acquisition. Some registries used data element collections specifically in case of established questionnaires. Some entries from registry E were tagged as calculated, representing entries that are not relevant for data acquisition but relevant for data storage. Registry F delivered an elaborated legacy format of its metadata offering consequently a short format of a data element like “date_birth” and a description like “date of birth”. In case of questionnaire data elements, the description included a direct question to the recording person. Furthermore, the description established links between two data elements by repeating the counterpart in square brackets.

Technical information was found in the registries’ deliveries too. Registries A, C, and F reported primary or foreign keys. Denominations from registry E included guiding notes related to data acquisition, e.g. a request to record more information in case of a certain selection. Overall, data elements were filled with the following types of entries:

- Generally understandable full or abbreviated terms
- Terms with context information, terms with counters, terms with guiding notes for data acquisition, terms tagged as calculated
- Data element collections with terms combining different data elements

- Shortcuts or labels
- Questionnaires or instruments, questions or prompts
- Primary or foreign keys

4. Discussion

The registries provided different interpretations of the concept "data element" independently from the use of the recommended template. Comparing the feedbacks of the registries with the reference system in figure 1, it becomes clear that different perspectives were used all having an own value in the management of metadata. The feedbacks comprised the perspectives of data storage and data acquisition additionally to the perspective of the information needed to answer research questions. Consequently, the catalog of data elements was also used for the listing of technical labels as well as for the listing of questions presented on CRFs. Therefore, this work extends the understanding of the challenge. Not only the "variety of ways in which similar data are represented" [12] has to be handled, but also the different perspectives indentified here.

The example of quality of life (QoL) should highlight the dependencies between the different perspectives. Patient reported outcome could be in the focus of a research question. Health-related QoL is defined as information needed for answering this research question. The EQ-5D-5L is selected as instrument to measure QoL and listed in the catalog of data elements. Implementing the EQ-5D-5L in a patient app requires the presentation of five questions with five categories each, along with a vertical line with a length of 20 centimeters. This lead to five further data elements - mobility, self-care, usual activities, pain/discomfort, anxiety/depression - to store the results of data acquisition. Furthermore, an element labelled VAS (visual analogue scale) with a numerical datatype is added for the visual analogue scale. The numerical value of QoL is calculated from the five dimensions of the EQ-5D-5L using a value set available for different nations [13]. This calculation requires three further data elements, two for the master data of the value set and one for the result that will be used in statistical analyses. The catalog of data elements is finally filled from all three perspectives regarding QoL.

The reference system must be extended to represent domain specific supplements as well as technical elements found in the registries' documents (cf. figure 2). Unfortunately, existing work does not support this complex system. For example, the Common Healthcare Data Interoperability Project focusses on the simplified model in figure 1 with a data acquisition and a data storage perspective only [6]. ISO/IEC 11179 defines a data element as "a unit of data that is considered in context to be indivisible" [7]. Taking the example of QoL, this definition supports both, the recording of EQ-5D-5L as data element with the score as value and the recording of its five dimensions as data element with the value set of five categories each. However, the term "data element" has not been adopted by the Clinical Data Interchange Standards Consortium (CDISC) standards (cf. http://www.cdisc.org/). We found items, variables and labels depending on the specific CDISC module. From the point of view of metadata science, the registries' interpretation of the term "data element" comprises descriptive, structural and technical aspects [14].

The FAIR guiding principle that "(meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation" [15] is theoretically relevant, but nowadays not fully satisfiable. The concept of a data element must be interpreted at least from the perspectives of research questions, data acquisition and data storage to

fully capture the life cycle of a registry. The variety of different interpretations of this concept in the metadata from the six registries in our study highlight this notion. We plan to make the metadata of the six registries publicly available to further stimulate the transformation of the reference system into metadata methodology.

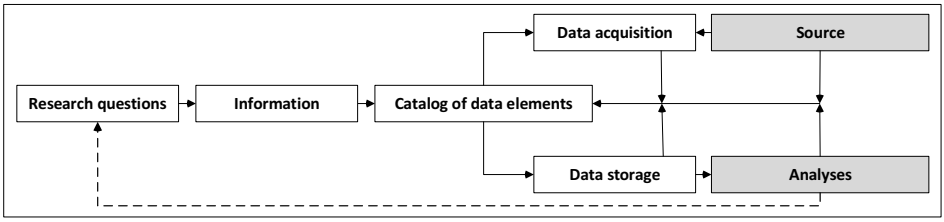


Figure 2. Extended reference system for metadata of registries.

Acknowledgements

Funded by the German Federal Ministry of Education and Research under contracts 01GY1901, 01GY1902, 01GY1903, 01GY1904, 01GY1905, 01GY1906, 01GY1917B.

References

[1] Irgens LM. The origin of registry-based medical research and care. Acta Neurol Scand. 2012;126 (Suppl 195):4-6.
[2] Baig, MA, Bazarbashi MS, AlDakhil H, Baig SM. The journey of clinical registries through various phases of the digital age: a technical perspective. Stud Health Technol Inform. 2022;289:345-8.
[3] Ulrich H, Kock-Schoppenhauer AK, Deppenwiese N, Gött R, Kern J, Lablans M, Majeed RW, Stöhr MR, Stausberg J, Varghese J, Dugas M, Ingenerf J. Understanding the nature of metadata: systematic review. J Med Internet Res. 2022;24:e25440.
[4] Glicklich RE, Leavy MB, Dreyer NA, eds. Registries for evaluating patient outcomes: a user's guide. 4th ed. AHRQ Publication No. 19(20)-EHC020. Rockville, MD: Agency for Healthcare Research and Quality; September 2020.
[5] Dugas M. Design of case report forms based on a public metadata registry: re-use of data elements to improve compatibility of data. Trials. 2016;17:566.
[6] Tcheng JE, Drozda JP, Gabriel D, Heath A, Wilgus RW, Williams M, Windle TA, Windle JR. Achieving Data Liquidity: Lessons Learned from Analysis of 38 Clinical Registries (The Duke-Pew Data Interoperability Project). AMIA Annu Symp Proc. 2020;2019:864-73.
[7] ISO/IEC 11179-3:2013(E), Information technology – Metadata registries (MDR). Part 3: registry metamodel and basic attributes, Third edition, 2013-02-15.
[8] Stausberg J, Harkener S, Semler S. Recent trends in patient registries for health services research. Methods of Information in Medicine. 2021; 60 (S 01):e1-e8.
[9] Stausberg J, Harkener S. Metadata of registries: results from an initiative in health services research. Stud Health Technol Inform. 2021; 281:18-22.
[10] Stausberg J, Harkener S. Bridging documentation and metadata standards: Experiences from a funding initiative for registries. Stud Health Technol Inform. 2019;264:1046-50.
[11] Leiner F, Haux R. Systematic planning of clinical documentation. Methods Inf Med. 1996; 35: 25-34.
[12] German E, Leibowitz A, Shahar Y. An architecture for linking medical decision-support applications to clinical databases and its evaluation. J Biomed Inform. 2009;42:203-18.
[13] Ludwig K, Graf von der Schulenburg JM, Greiner W. German Value Set for the EQ-5D-5L. Pharmacoeconomics. 2018;36:663-74.
[14] Riley J. Understanding metadata. What is metadata, and what is it for? Baltimore, MD: NISO, 2017.
[15] Wilkinson MD, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A, et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data. 2016;3:160018.

Challenges of Trustable AI and Added-Value on Health
B. Séroussi et al. (Eds.)

doi:10.3233/SHTI220433

Automatic Objective Severity Grading of Peripheral Facial Palsy Using 3D Radial Curves Extracted from Point Clouds

Tim BÜCHNER [a,1], Sven SICKERT [a], Gerd Fabian VOLK [b], Orlando GUNTINAS-LICHIUS [b], Joachim DENZLER [a]
[a] *Friedrich Schiller University Jena, Faculty of Mathematics and Computer Science, Computer Vision Group, Germany*
[b] *Jena University Hospital, Department of Otorhinolaryngology, Friedrich Schiller University Jena, Germany*

Abstract. Peripheral facial palsy is an illness in which a one-sided ipsilateral paralysis of the facial muscles occurs due to nerve damage. Medical experts utilize visual severity grading methods to estimate this damage. Our algorithm-based method provides an objective grading using 3D point clouds. We extract from static 3D recordings facial radial curves to measure volumetric differences between both sides of the face. We analyze five patients with chronic complete peripheral facial palsy to evaluate our method by comparing changes over several recording sessions. We show that our proposed method allows an objective assessment official palsy.

Keywords. Facial Palsy, Point Clouds, Grading

1. Introduction

Peripheral facial palsy originates from underlying damage to the nervus facialis by infections, brain tumors, strokes, or injuries during surgery. The damage to the nerve can vary, resulting in mild to severe symptoms. A comparison between paresis and healthy side or structural changes are the primary evaluation criteria in grading systems like Sunnybrook [1], House-Brackmann [2], or eFace [3]. Such grading systems are prone to a subjective bias introduced by the rater's grading experience and the evaluation environment as stated in [4].

A computer-aided approach [5] uses 2D facial landmarks for precise measurements. However, algorithms for landmark estimation do not include facial palsy patient images in their training data. Thus, placing is unreliable for medical evaluation as such cases are unknown to the algorithm. Deep learning approaches [6,7] recreate existing metrics from 2D images and copy human grading biases. Furthermore, essential surface structures, like the nasolabial fold depth, cannot be reliably estimated from 2D recordings. The use of 3D facial data is promising for fine-grained assessments. One approach [8] proposes sparse 3D landmarks features, which could be used to describe anteroposterior direction changes. However, current applications require hand-operated processes for correct

[1] Corresponding Author: Tim Büchner, Friedrich Schiller University Jena, Ernst-Abbe-Platz 2, 07743 Jena, Germany; E-mail: tim.buechner@uni-jena.de.

execution. In contrast, our method automatically creates a dense, structured 3D facial surface description based on a 3D camera recording. Our method's algorithmic nature decreases the potential for subjective biases as we exclusively use the 3D sensor data.

2. Materials and Methods

Facial radial curves were first proposed by Berretti et al. [9] using geodesic lines as novel face representation. They are used to describe anatomical structures [10] or to achieve presentation attack detection [11]. We provide a novel extraction algorithm for 3D point clouds creating equidistant sampled curves around the nose tip, see Figure 1. These curves allow us to objectively measure the volumetric differences between both sides of the face and should indicate facial nerve damage.

2.1. Acquisition and Pre-processing of the 3D Point Clouds

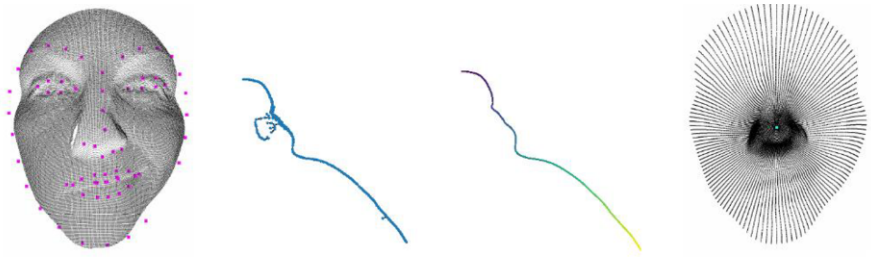
To acquire the data for evaluation, we use a 3D sensor, which provides a color and depth image. The sensor captures a patient's frontal view while following a set of instructions [4] showing different facial movements. The method described in [11] allows us to compute a 3D point cloud including 68 facial landmarks detected in the image. We translate the coordinate system origin to the nose tip for simplified rotation and cropping of the point cloud. We define a cropping sphere with a radius from the nose to the jaw tip and remove all points outside it. This step ensures that unnecessary points, e.g., hair or upper body, are removed. Estimating the facial radial curves requires that the nose points towards the positive *z*-axis. We calculate the required rotation with the method described in [11] using 3D landmarks.

2.2. Robust Estimation of Facial Radial Curves in 3D

We estimate N facial radial curves rotating around the nose tip in the *yz*-plane. Each curve describes a path along the face surface originating at the nose tip and ending on the outer radius of the face. A plane R defined by its normal vectors (n_1, n_2) is rotated around this *yz*-plane in *360/N* steps. We define the set of possible curve points as

$$P = \{p_i = (x_i, y_i, z_i) : |p_i^T n_1| \leq \delta;\ p_i^T n_2 > 0\}, \quad (1)$$

where δ is a threshold to ensure small distances to the curve location. Further, we apply a perpendicular transformation for each point in P such that they lie on the plane R. As all points reside on the same 2D plane, we can efficiently solve the problem in the two-dimensional space. As most points were correctly recorded by the sensor, there must exists a subset of points in P, closely describing the facial surface. We propose an extraction algorithm for this path that is resilient towards outliers, e.g., errors during the recording. First, we define a complete graph G within P. Then we compute the minimum spanning tree (*MST*) of G. This process reduces the number of possible paths drastically but ensures consideration of all points during the reconstruction. We extract the path from the nose tip to the outer radius point in the *MST*, and only a single path should exist in G. Outliers cannot be included as they connect to the graph via one edge.



(a) Facial point cloud with estimated landmarks (b) Min. spanning tree inside surface point set (c) Surface Reconstruction (d) Extracted facial radial curves

Figure 1. Facial curves intermediate construction steps

Inside the selected path, we fit a two-dimensional spline. We sample the spline with M equidistant steps to create a discrete curve representation *C*. The resulting curve points are placed back in the plane *R* to ensure the correct three-dimensional rotation. After the computation of all curves, we obtained a dense, structured representation of the face in $\mathbb{R}^{M\times N\times 3}$, see Figure 1 (d).

2.3. Grading Peripheral Facial Palsy using Facial Radial Curves

To estimate the severity of facial palsy, we compare both sides of the face using the computed facial radial curves. We define corresponding curves as the index pair $(i, N - i)$ with $i = 1 \ldots N/2$. They share the exact origin, and each point pair describes the same face location. We can use this property to compute the volume difference, see Eq. (2), in the anteroposterior direction between the curves, which is the mean absolute pointwise ($\ominus$) volume difference of the z-coordinate:

$$d_i = \emptyset\left(\left|C_z^{(i)} \ominus C_z^{(N-i)}\right|\right). \tag{2}$$

We further provide a mapping *m()*, see Eq. (3), to simplify usage similar to Sunnybrook [1] with 100 being a perfect symmetric face and 0 a complete asymmetry:

$$m(d_i) = 100 - \frac{arctan(d_i)}{2/\pi} \times 100. \tag{3}$$

We repeat this process for all curve pairs to obtain an average value for the recording. Symmetric faces would yield a score of 100, and values lower than 100 indicate some asymmetry. However, every human face has some dissymmetry. It is essential to measure unaffected faces to obtain a threshold to understand changes due to facial palsy.

3. Results

We compare t = 3 time points of five long-term patients with chronic complete peripheral facial palsy for two years. For each recording, we extract the facial radial curves with parameters $N = M = 128$. In Table 1 we list for each patient our proposed grading and the average Sunnybrook grading of three independent medical experts. For our grading,

we display the average scores of all movements. Furthermore, we use the standard deviation to indicate symmetry among the movements. A small value would indicate high symmetry among all movements, whereas a high value would indicate strong asymmetry. We can see an improvement of patients one to four in terms of the average score, including a decreased standard deviation. However, the results of patient five indicate no symmetry improvement with consistent variance among all movements at the first and most recent session. These improvements correspond with the observations of the medical experts who treat these patients based on their grading scores. However, the deviations for the Sunnybrook score elevate the problem of human bias in the evaluation of facial palsy.

Table 1. Scores of chronic complete peripheral facial palsy patients for two years. Our method shows average score and standard deviation of eleven movements [4]. For Sunnybrook, the average score and standard deviation of three independent raters are shown. Scores written in italic represent incomplete scores.

Patient	Grading	T1	T2	T3
1	Volume Difference	76.56 ± 8.44	83.37 ± 5.47	84.09 ± 3.71
	Sunnybrook	17.67 ± 5.03	28.00 ± 4.36	41.50 ± 6.36
2	Volume Difference	73.51 ± 9.36	76.47 ± 10.66	82.73 ± 5.90
	Sunnybrook	12.33 ± 3.51	25.67 ± 3.06	20.00 ± 3.61
3	Volume Difference	66.44 ± 10.32	71.83 ± 10.81	71.56 ± 7.91
	Sunnybrook	*12.50 ± 2.50*	16.66 ± 2.88	22.67 ± 2.08
4	Volume Difference	66.84 ± 9.42	74.22 ± 6.98	71.92 ± 7.41
	Sunnybrook	*17.50 ± 2.50*	16.33 ± 3.06	18.67 ± 4.93
5	Volume Difference	71.26 ± 10.05	74.92 ± 6.70	70.68 ± 11.77
	Sunnybrook	*13.33 ± 2.88*	24.00 ± 6.24	29.67 ± 11.37

4. Discussion

The results in Table 1 expound that our proposed method is a suitable alternative for grading facial palsy. A significant problem for existing gradings is the variance between raters which obstructs an objective assessment. Our data-driven method requires only a 3D recording of the patient doing different movements. With precise instructions, this process can be completely automated and reduce the time constraints for medical experts. The usage of 3D data further elevates the objective natures of our grading. Using a large set of different facial movements ensures a better understanding of the impact of the nerve damage. Additionally, tracking inter-patient scores might be an indicator of successful treatment.

5. Conclusion

Our proposed data-driven method yields objective, comparable results to existing gradings in the first experiments. The recording process can be fully automated with a photo laboratory and offers medical experts efficient and fast feedback. We plan to

evaluate our method on a more extensive and diverse set of patients during the clinical routine. However, our first results indicate that it could be an additional routine during the assessment of facial palsy. We intend to continue this research track by including more dynamic recordings of the patients' facial movements. Lastly, approaches to separate the curves into different regions, like House-Brackmann [2], could further enhance the grading.

References

[1] Ross BG, Fradet G, Nedzelski JM. Development of a Sensitive Clinical Facial Grading System. Otolaryngol Head Neck Surg. 1996 Mar;114(3):380-6.
[2] House JW, Brackmann DE. Facial Nerve Grading System. Otolaryngology–Head and Neck Surgery: Official Journal of American Academy of Otolaryngology–Head and Neck Surgery. 1985 Apr;93(2): 146-7.
[3] Banks CA, Bhama PK, Park J, Hadlock CR, Hadlock TA. Clinician-Graded Electronic Facial Paralysis Assessment: The eFACE. Plast Reconstr Surg. 2015 Aug;136(2):223e.
[4] Volk GF, Schaede RA, Thielker J, Modersohn L, Mothes O, Nduka CC, et al. Reliability of grading of facial palsy using a video tutorial with synchronous video recording. Laryngoscope. 2019;129(10).
[5] Gaber A, Taher MF, Wahed MA. Quantifying Facial Paralysis Using the Kinect V2. Conf Proc IEEE Eng Med Biol Soc. 2015; 2015:2497-501.
[6] Raj A, Mothes O, Sickert S, Volk GF, Guntinas-Lichius O, Denzler J. Automatic and Objective Facial Palsy Grading Index Prediction using Deep Feature Regression. Annual Conference on Medical Image Understanding and Analysis (MIUA). 2020:253-66.
[7] Sajid M, Shafique T, Baig MJA, Riaz I, Amin S, Manzoor S. Automatic Grading of Palsy Using Asymmetrical Facial Features: A Study Complemented by New Solutions. Symmetry. 2018;10(7).
[8] Özsoy U, Uysal H, Hizay A, Sekerci R, Yildirim Y. Three-Dimensional Objective Evaluation of Facial Palsy and Follow-up of Recovery with a Handheld Scanner. J Plast ReconstrAesthet Surg. 2021 Jun
[9] Berretti S, Bimbo AD, Pala P, Mata FJS. Face Recognition by SVMs Classification and Manifold Learning of 2D and 3D Radial Geodesic Distances. Eurographics 2008 Workshop on 3D Object Retrieval. 2008:00.
[10] Bowman AW, Katina S, Smith J, Brown D. Anatomical Curve Identification. Comput Stat Data Anal. 2015; 86:52-64.
[11] Thümmel M, Sickert S, Denzler J. Facial Behavior Analysis using 4D Curvature Statistics for Presentation Attack Detection. IWBF. 2021:1-6

Challenges of Trustable AI and Added-Value on Health
B. Séroussi et al. (Eds.)

doi:10.3233/SHTI220434

Linking EMR Data to REDCap: Implementation in the SOLKID Register

Tobias J. BRIX[a,1,2], Leonard GREULICH[a,1], Alice JANSSEN[a], Sarah RIEPENHAUSEN[a], Philipp NEUHAUS[a], Johannes OEHM[a], Jeannine WEGNER[b], Barbara SUWELACK[b], Michael STORCK[a,1] and Julian VARGHESE[a,1]
[a] *Institute of Medical Informatics, University of Münster, Germany*
[b] *Department of Medicine D, Transplantnephrology, University of Münster, Germany*

Abstract. Secondary use, the reuse of medical patient data stored during routine care in the hospital's electronic medical records (EMR) for research purpose is common, especially for registers and pragmatic trials. Often the medical data items are copied manually from the EMR into the used research database. This process is time consuming and error prone. In the "Safety of the Living Kidney Donor – The German National Register" (SOLKID-GNR), laboratory results gathered during control check-ups of the living donors before and after the transplantation are to be transferred from the EMR into the electronic data capture system REDCap of the register. In this work, we present our approach of realizing an automated transfer of time-dependent laboratory results from the EMR of the University Hospital of Münster to REDCap. A challenge lies in the multi-center structure of SOLKID-GNR. The participating transplant centers are using different EMR systems, which requires a flexible architecture design. In addition, we aimed to support reuse of the implementation for other research settings with other medical data items of interest.

Keywords. Secondary use, SOLKID-GNR, REDCap

1. Introduction

Medical patient data collected during routine care in hospitals is stored in electronic medical records (EMR). In contrast, medical patient data for scientific research purposes, e.g., clinical trials, is collected in specific electronic data capture systems (EDC). Many data items can be directly reused from routine care. This is a common approach known as secondary use and is often applied for registers and pragmatic trials. Thus, for instance, laboratory observations do not have to be performed twice, which saves costs and labor. However, a transcription from the EMR to an EDC is often time consuming and error prone, since manual copying of information always implies the risk of human mistakes. Therefore, an automated transfer is preferable.

The German living donor register "Safety of the Living Kidney Donor – The German National Register" (SOLKID-GNR) collects data of the medical and psychosocial outcome of living kidney donors [1]. As part of the register, donors are

[1] These Authors contributed equally.

[2] Corresponding Author, Tobias J. Brix, Institute of Medical Informatics, University of Münster, Albert-Schweitzer-Campus 1, Building A11, 48149 Münster, Germany; E-mail: tobias.brix@uni-muenster.de.

monitored before and after the transplantation in multiple check-ups. During these check-ups, laboratory observations are documented, e.g. creatinine. Since living donors must be monitored very strictly in the time after transplantation anyway, many data items, especially laboratory results, are already documented during routine care. Previously, this given data was transferred manually by study nurses for secondary use. SOLKID-GNR is a multi-center register consisting of multiple hospitals. While the destination of an automated data transfer is always the same EDC system, the source of the data transfer varies. Each hospital is using its own EMR, and even if the same EMR is used, these systems can be configured individually and the same data can be stored in different locations and formats.

Objective of this work is the implementation of a prototype, which allows an automated transfer of medical patient data from the EMRs of different healthcare centers to the EDC of SOLKID-GNR. The current prototype is focused on the EMR of the University Hospital of Münster (UKM) and the current use case is the SOLKID-GNR register. The implementation is aimed to be as generic as possible by utilizing technical and semantical standards. This allows reuse in future research projects and different EMR installations.

2. Methods

2.1. SOLKID-GNR Setting and Challenges

The German living donor register SOLKID-GNR consists of 38 German transplant centers. A single "Research Electronic Data Capture" (REDCap) instance is used as central EDC system storing the data of all 38 centers [2]. Data access groups (DAG) of REDCap are used to ensure that a center has only access to its own patients. The pseudonymization of the data records is carried out by a central pseudonymization service "Mainzelliste" [3], which is seamlessly connected to REDCap [4].

The first challenge is the diversity of EMRs used by each transplant center. This requires an individual customizable middleware to connect to the specific EMRs. If a connection to the EMR is established, the second challenge is the identification of the correct patient. REDCap only stores the pseudonym of a patient while the EMR identifies patients by their identifying data or special IDs. Third, the correct data within the EMR must be identified. Finally yet importantly, a time component must be considered. SOLKID-GNR has longitudinal follow-ups and the used data must be associated to the given time period.

2.2. Implementation Details

While REDCap can be used without cost under the REDCap license, it is no open-source software and does not allow complete customization. However, REDCap supports a powerful extension and plugin mechanism based on modules. These modules can be implemented in PHP for the back-end and HTML5, CSS and JavaScript in the front-end. In this paper, we have implemented a specific REDCap module for our use case.

As middleware, i.e., connection service between REDCap and the EMR, we use a RESTful web-service implemented in the Java Spring Framework. This middleware was already in use, but had to be extended to support the current workflow [5]. The current implementation is limited to the UKM, which is using Dedalus ORBIS as EMR.

Specifying the exact semantic meaning of a medical data item is a common problem in medical informatics. The current state of the art is to use codes of classifications and terminologies to be independent of language, spelling and formulation. In our use case of the transfer of laboratory results, the "Logical Observation Identifiers Names and Codes" (LOINC) is most relevant, a terminology for laboratory and clinical observations.

3. Results

To use the implemented REDCap module, it must be installed on the REDCap server and be activated for the project. Afterwards, the module must be configured by selecting data items on "electronic case report forms" (eCRF) that should be queried. On selected forms, a new button is rendered in the normal data entry mode, which triggers the data transfer. Since the currently opened form is associated with a patient record, REDCap knows the record's pseudonym and can send it to the middleware. Additionally, an individual mapping between LOINC codes and each selected data item of the eCRF must be configured. Thus, during data request, REDCap knows the associated codes of the opened form and can send the list to the middleware. To achieve the time dependency, an interval can be specified for each form. Since values like, one week before or two weeks after the follow up cannot be specified in advance, this configuration can be done relatively to a reference date element. In the case of SOLKID-GNR, an item in the eCRF contains the date of follow up and the interval sent to the middleware is calculated relatively to this reference date.

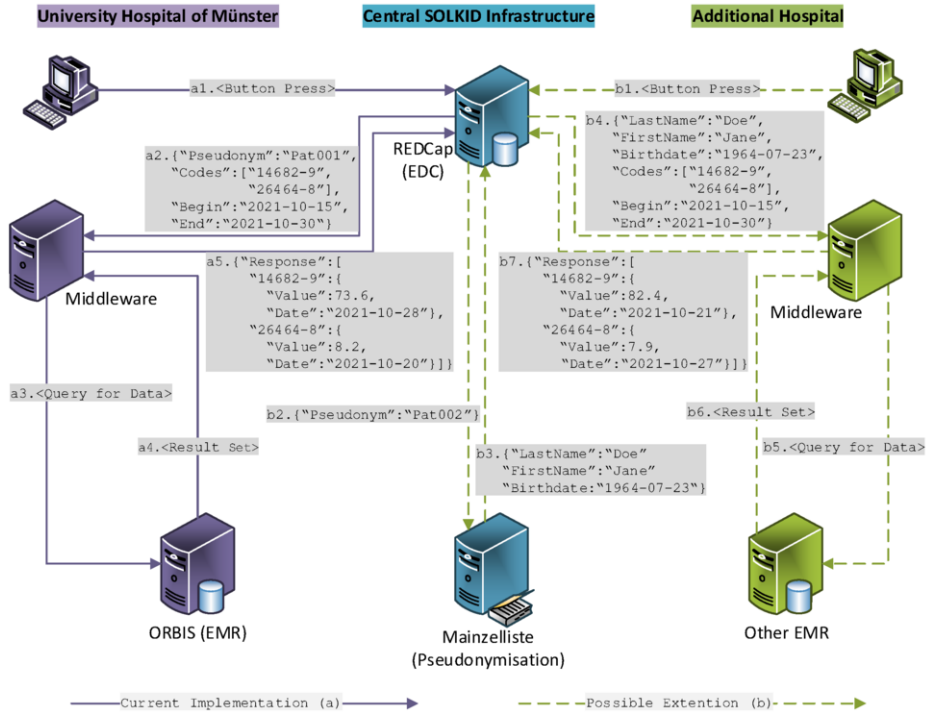


Figure 1. Architecture and workflow of the secondary use in SOLKID-GNR.

The currently implemented architecture and workflow can be seen on the left side of Figure 1. If a user triggers the data transfer in REDCap (1.), the EDC communicates with the middleware's REST-API via an authenticated HTTPS request and submits a patient's pseudonym, the data elements of interest coded as a list of LOINC codes and a time interval that constrains the laboratory observation's documentation date in the EMR (2.).

Once the request is received by the middleware, it checks whether the given pseudonym is registered in ORBIS for SOLKID-GNR. In Münster, an EMR-integrated system is in use, which allows the management of research projects and the registration of their participants [6]. The participant's pseudonym is part of the registered properties. If a patient is not yet registered, a HTTP-error code is returned and a meaningful error message is displayed in REDCap. If the pseudonym and the associated patient have been identified successfully, the data export from the EMR begins (3.). Based on the submitted time interval, the associated medical data is extracted. If multiple laboratory results are found in the given interval, the newest observation is returned (4.). At the moment, the mapping from LOINC codes to laboratory observation IDs in ORBIS are based on a manual mapping file provided by the local medical data integration center. A map, containing the LOINC codes as keys and the found laboratory results and associated observation dates, is sent back to REDCap (5.).

During data retrieval, the previous module configuration is used to populate the results into the correct eCRF data item in REDCap. Although optional, we have extended the eCRF by a laboratory date item for each laboratory result. This is needed since laboratory observations may have been determined on different dates. To ensure reproducibility and traceability, these exact dates have to be documented. As optimization, the most common date is calculated by the module and stored as primary date, and only deviating dates are documented or hidden otherwise. This keeps the eCRF clear if all values are from a single observation, but also provides the flexibility of multiple observations.

4. Discussion

The presented approach is working successfully for the site Münster of the SOLKID-GNR. However, it has its limitations regarding the transition to other participating hospitals and to other research projects. While the REDCap module can be configured individually for each project and can be adapted for each new eCRF, the middleware, i.e., the export process itself, is highly EMR and medical concept dependent. Each change requires an individual adaptation. Even hospitals also using ORBIS may not be able to use the implementation without adaptation, since ORBIS internal structures are customizable. It is worth mentioning that a communication server may replace the middleware at some hospitals. We used the existing middleware since it was easier to configure and adapt.

Currently, we only applied LOINC in our approach. If more data besides laboratory or clinical observations should be identified, possible options would be "International Statistical Classification of Diseases and Related Health Problems" (ICD) for diagnoses, "Systematized Nomenclature of Medicine – Clinical Terms" (SNOMED-CT) as the reference terminology for medical concepts, or the "Unified Medical Language System" (UMLS), a metathesaurus combining all the previously mentioned ones. All these systems are using fixed codes, i.e., strings, to encode medical concepts. The APIs of both

REDCap and middleware can handle any string-based codes, and therefore any code system best suited for the given application case can be used. Only the mapping from codes to EMR elements must be adapted in the middleware. This mapping is EMR and data item dependent.

Another limitation is the identification of a patient in the EMR only knowing the pseudonym of the register. The beneficial situation of having the pseudonyms present in the EMR is often not given in other hospitals. A possible solution is illustrated on the right side of Figure 1. A patient in the EMR can only be identified by his or her identifying data like first name, last name and birthdate. The Mainzelliste allows a re-identification based on the pseudonym. Thus, the identifying data can be automatically fetched by depseudonymization and sent to the middleware to find and identify the associated patient. This approach rises additional data protection and ethical concerns, which must be addressed. Furthermore, the general issue about how to uniquely identify a person has to be discussed, since these three attributes do not have to be distinctive. The Mainzelliste also allows custom attributes. Storing the EMR-ID or social security number during the pseudonymization process could be a solution.

5. Conclusions

In this work, we presented our implementation of the SOLKID-GNR register at the site of Münster to support the automated transfer of time dependent laboratory results from the EMR to the research database REDCap. While the approach is feasible for a single site, its reuse for other sites and research projects requires additional work to establish site-specific interfaces for querying the respective EMRs. The challenges and possible solutions for this adaptation have been discussed.

6. Acknowledgements

Supported by BMBF grant Ns. 01ZZ1802V (HiGHmed/Münster) and 01GY1906 (SOLKID-GNR/Münster). We would like to thank the entire SOLKID team.

References

[1] Suwelack B, Dugas M, Koch M, Sommerer C, Urban M, Gerß J et al. Die Sicherheit des Lebendnierenspenders – Das Deutsche Lebendspende Register SOLKID-GNR – Entstehung und Struktur eines nationalen Registers in der Versorgungsforschung. Gesundheitswesen (Bundesverband der Arzte des Offentlichen Gesundheitsdienstes (Germany)) 2021;83(S 01):S33-S38.

[2] Harris PA, Taylor R, Thielke R, Payne J, Gonzalez N, Conde JG. Research electronic data capture (REDCap)--a metadata-driven methodology and workflow process for providing translational research informatics support. Journal of biomedical informatics 2009;42(2):377–81.

[3] Lablans M, Borg A, Ückert F. A RESTful interface to pseudonymization services in modern web applications. BMC medical informatics and decision making 2015;15:2.

[4] Greulich L, Brix TJ, Storck M, Dugas M. A Seamless Pseudonymization and Randomization Workflow for REDCap. Studies in health technology and informatics 2021;281:952–6.

[5] Bruland P, Doods J, Brix T, Dugas M, Storck M. Connecting healthcare and clinical research: Workflow optimizations through seamless integration of EHR, pseudonymization services and EDC systems. International journal of medical informatics 2018;119:103–8.

[6] Brix T, Gessner S, Bruland P, Dugas M. CIS-integrated study patient management: registration, pseudonymization and randomization: German Medical Science GMS Publishing House; 2018.

Challenges of Trustable AI and Added-Value on Health
B. Séroussi et al. (Eds.)

doi:10.3233/SHTI220435

Evaluation of a Personal Digital Assistant Device Implementation for Barcode Medication Administration with Nurses Using a Likert Questionnaire

Juan DESCALZO[a,1], Javier CASTRO[a], María SMITH[a] and Daniel LUNA[a]
[a] *Health Informatics Department, Hospital Italiano de Buenos Aires*

Abstract. The majority of preventable medication errors occur at the administration stage. BCMA systems are used to improve safety and prevent errors in this stage. A variety of devices can be used for this purpose. Hospital Italiano de Buenos Aires is a high complexity medical center implementing a BCMA project since 2015. It is currently migrating to PDA devices for nurses. The objective of this work is to evaluate the implementation of these new devices in selected wards at HIBA using a self-reported questionnaire. From 318 contacted nurses, 58 answered the questionnaire (18.2% response rate). Overall, agreement was high among all statements regarding the new devices. Nurses valued especially the increased safety to reduce errors, improvements in previous hospital processes and achieving improvements in the flow and quality of patient care. Nurses recommended the use of the device in their sector, with a mean score of 4.6/5 and 91.3% agreement, highest in total. This proved to be a cost-effective method of evaluation of the newly implemented devices and acceptance by nurses. Measures to incorporate the remaining nurses' feedback should be considered.

Keywords. BCMA, nursing, implementation evaluation

1. Introduction

The majority of preventable medication errors at the inpatient setting occur in the administration stage [1], as many as 34% according to Bates et al [2]. The Bar-Coded Medication Administration (BCMA) system makes the process safer by reading a barcode on the patient's bracelet and on medication pouches to ensure the 5 rights of medication safety: medication, patient, time, dose and route [3]. Personal Digital Assistants (PDAs) are small and portable handheld computers [4]. PDAs with a barcode scanner, linked to the hospital's computer system, have been used to ensure safety and reduce errors in hospitalized patients [5]. The Hospital Italiano de Buenos Aires (HIBA) implemented BCMA in 2015, but over the years, new technologies such as PDAs have been added. Suitable equipment selection for this project is considered a key to success.

[1] Corresponding Author, Juan Descalzo, Health Informatics Department, Hospital Italiano de Buenos Aires, Tte. Gral. Juan D. Perón 4190, C1199ABB Buenos Aires, Argentina; E-mail: juan.descalzo@hospitalitaliano.org.ar.

The objective of this work is to evaluate the implementation of PDA devices for Bar-Coded Medication Administration in some general wards at the Hospital Italiano de Buenos Aires (HIBA) using a self-reported Likert-type scale questionnaire.

2. Methods

2.1. Setting

The Hospital Italiano is a high complexity medical center located in the City of Buenos Aires. It has 750 inpatient beds, 200 for critical care. It employs more than 9000 people, including 1600 nurses. It is a university hospital that covers the entire spectrum of health care, performing around 30000 discharges per year. Since 1998, HIBA has had its own health information system [6] and it's a Stage 7 HIMSS certified institution.

A BCMA system has been implemented at the hospital since 2015, on desktop devices [7]. Starting 2017 mobile devices have been the main technology used, such as tablets incorporated into "Workstation on Wheels" or wall-mounted. Also, nurses have the possibility of using the same mobile application on their personal cell phone [8].

At the time of this work, the Health Informatics Department is going forward with the implementation of Honeywell EDA51 PDA devices for its BCMA project.

2.2. Questionnaire

A Likert scale [9] questionnaire using the Google Forms platform. The survey consisted of 14 statements regarding the recently implemented devices. Nurses could answer on a 5-point scale: 1 = Strongly Disagree, 2 = Disagree, 3 = Undecided or Neutral, 4 = Agree, and 5 = Strongly Agree. Questions were designed using a general IT implementation questionnaire already in use in HIBA, selecting and adapting questions for these study objectives, with support from research methodology specialists.

2.3. Sample

The implementation team selected 7 wards to survey: 3 general adult, 2 general pediatric, 1 orthopedic and 1 obstetrics ward. All of these wards were part of the implementation plan of the new PDA devices, jointly chosen with the Nursing Department. Supervisors were informed of the survey and accepted to participate.

All nurses following these inclusion criteria were contacted:

- Regular nursing staff from the selected wards in January 2021.
- Not a supervisor or supervisor aid.
- Not on vacation or other type of leave during January or February 2021.

In total, the research team contacted 318 nurses via email at their work address. Participation was voluntary and confidential

3. Results

Sample characteristics are summarized in Table 1 and the questionnaire responses in Table 2, a numerical score was assigned to each answer (from 1 for "Strongly Disagree" to 5 for "Strongly Agree") and a Mean Score was calculated for each statement..

Table 1. Sample characteristics.

Characteristic	Value
n	58
Female (%)	49 (84.4)
Ward	
General Adult (%)	40 (68.9)
General Pediatric (%)	6 (10.3)
Orthopedics (%)	7 (12)
Obstetrics (%)	5 (8.6)
Shift	
Morning (%)	12 (20.7)
Afternoon (%)	10 (17.2)
Night (%)	17 (29.3)
Weekends (%)	19 (32.8)

From 318 contacted nurses, 58 answered the questionnaire (18.2% response rate). General Adult wards are more represented (68.9%) than others. This was expected as half they represented both half the wards surveyed and the highest staffed.

Table 2. Summary of responses.

Statement	Strongly Disagree (1)	Disagree (2)	Undecided or Neutral (3)	Agree (4)	Strongly Agree (5)	Mean Score
Using the new device reduces errors.	1 (1.7%)	0 (0%)	7 (12.1%)	17 (29.3%)	33 (56.9%)	4.3
The use of the new device has achieved greater adherence to politics and procedures.	0 (0%)	3 (5.2%)	8 (13.8%)	18 (31%)	29 (50%)	4.3
Problems with the new device don't interfere with the patient care process.	6 (10.3%)	4 (6.9%)	14 (24.1%)	5 (8.6%)	29 (50%)	3.8
I feel able to assist others in using the new device.	1 (1.7%)	3 (5.2%)	3 (5.2%)	7 (12.1%)	44 (75.9%)	4.6
Using the new device is more efficient than doing things the previous way	3 (5.2%)	4 (6.9%)	2 (3.4%)	7 (12.1%)	44 (75.9%)	4.4
Using the new device allows me to spend more time on some other aspects of the care process.	5 (8.6%)	3 (5.2%)	13 (22.4%)	12 (20.7%)	25 (43.1%)	3.8
I believe that the use of the new device improves the quality of patient care.	2 (3.4%)	5 (8.6%)	4 (6.9%)	12 (20.7%)	35 (60.3%)	4.3

The information provided by the new device allows me to make better decisions about patient care.	2 (4.7%)	4 (9.3%)	10 (23.3%)	9 (20.9%)	18 (41.9%)	3.8
The implementation of the new device takes into account the specific needs of my service area.	2 (3.4%	4 (6.9%)	18 (31%)	14 (24.1%)	20 (34.5%)	3.8
The people I work with daily help me with the use of the new device.	11 (19%)	3 (5.2.%)	16 (27.6%)	8 (13.8%)	20 (34.5%)	3.4
In general, I prefer to use the new device rather than to do things the way they were done before.	1 (1.7%)	3 (5.2%)	9 (15.5%)	12 (20.7%)	33 (56.9%)	4.3
The PDA device is practical in size and weight for my work.	1 (1.7%)	3 (5.2%)	7 (12.1%)	12 (20.7%)	35 (60.3%)	4.3
Code reading (patient bracelet, drugs) with the device is easy and effective.	1 (1.7%)	2 (3.4%)	5 (8.6%)	11 (19%)	39 (67.2%)	4.5
I would recommend the use of this device in my sector.	1 (1.7%)	0 (0%)	4 (6.9%)	14 (24.1%)	39 (67.2%)	4.6

In general, there was high agreement on all statements. The statements "I feel able to assist others in using the new device" and "Using the new device is more efficient than doing things the previous way" presented the highest "Strongly Agree" scores (75.9%).

The highest disagreement was seen on the statements "Problems with the new device don't interfere with the patient care process" (10.3% Strongly Disagree, 6.9% Disagree) and "The people I work with daily help me with the use of the new device" (19% Strongly Disagree, 5.2% Disagree).

On the final recommendation to use the device in their sector, agreement was high, with a mean score of 4.6 and 91.3% some degree of agreement, highest in total.

4. Discussion

The results showed that the nurses agreed with the implementation based on increased safety to reduce errors in drug administration, better adherence to policies, improvements in previous hospital processes, achieving improvements in workflow and patient care quality.

Although the nurses felt capable of helping others with the use of PDAs, there was a low report of collaboration between them. The implementation team should continue to work on education and generate teamwork strategies between them.

This study has some limitations. Nurses may have been biased to show a high degree of agreement due to the possibility of receiving more PDAs in their ward. Finally, among the wards consulted, general adult sectors were over represented. Emergency or critical care areas should be taken into account in following studies. While response rate was

average for this type of study, measures to reach the remaining nurses should be considered. We planned to share with them this paper results asking for further feedback and collaborate with supervisors to increase the response rates.

This type of low-cost intervention can be used to evaluate implementation and generate collaboration in IT projects directly with end-users and in their work environment. In this case, a direct line of work with nurses, a sector underrepresented in the general literature and often overlooked by management.

5. Conclusions

The objective of this study was to measure nurses' agreement towards implementation of PDA devices using a self-reported questionnaire. The team achieved good collaboration with nurses. Response rate was acceptable for the intent of this work.

Nurses showed general agreement with all statements, suggesting PDAs are a good device choice for this particular implementation. Recommendation of use of these technologies was very high among nurses. This information is highly valuable for management to make informed decisions and device acquisition. This methodology was low-cost for the implementation team and proved useful for project monitoring.

References

[1] Paoletti RD, Suess TM, Lesko MG, et al. Using bar-code technology and medication observation methodology for safer medication administration. Am J Health Syst Pharm. 2007;64(5):536–543.
[2] Bates DW, Cullen DJ, Laird N et al. Incidence of adverse drug events and potential adverse drug events: implications for prevention. JAMA. 1995; 274:29-34,
[3] van Onzenoort HA, van de Plas A, Kessels AG, Veldhorst-Janssen NM, van der Kuy PH, Neef C. Factors influencing bar-code verification by nurses during medication administration in a Dutch hospital. Am J Health Syst Pharm. 2008 Apr 1;65(7):644-8.
[4] Puskar KR, Aubrecht J, Beamer K, Carozza LJ. (2004). Implementing information technology in a behavioral health setting. Issues in Mental Health Nursing, 25(5), 439-450.
[5] Baumgart DC. Personal digital assistants in health care: experienced clinicians in the palm of your hand? Lancet. 2005 Oct 1;366(9492):1210-22. doi: 10.1016/S0140-6736(05)67484-3.
[6] Plazzotta F, Luna D, González Bernaldo de Quirós F. (2015). Sistemas de información en salud: integrando datos clínicos en diferentes escenarios y usuarios. Revista Peruana de Medicina Experimental y Salud Pública, 32(2), 343-351.
[7] Maydana T, Giraldo L, González ZA, Schachner B, John III C, Luna DR, et al. (2017, January). BarCode Medication Administration in ICU: Learning from Our Nurses. In MedInfo (pp. 1095-1098).
[8] Giraldo L, Schachner B, Luna D, Benítez S. (2018). Exploring Nurses' Perceptions and Expectations Toward a BCMA Implementation Using a Mobile App and Workstations as a Change Management Strategy. Studies in health technology and informatics, 250, 134-138.
[9] Likert R. (1932). A technique for the measurement of attitudes. Archives of psychology.

Challenges of Trustable AI and Added-Value on Health
B. Séroussi et al. (Eds.)

doi:10.3233/SHTI220436

Federated Learning in Healthcare: A Privacy Preserving Approach

K NARMADHA[1] and P VARALAKSHMI
Department of Computer Technology, Anna University, MIT Campus, Chennai, India

Abstract. A need to enhance healthcare sector amidst pandemic arises. Many technological developments in Artificial Intelligence (AI) are being constantly leveraged in different fields of healthcare. One such advancement, Federated Learning(FL) has acquired recognition primarily due to its decentralized, collaborative nature of building AI models. The most significant feature in FL is that, raw data remain with the data sources throughout the training process and thus preventing its exposure. Hence, FL is more suitable and inevitable in healthcare domain as it deals with private sensitive data which needs to be protected. However, privacy threats still exist in FL, necessitating a requirement for further improvement in privacy protection This paper discusses about the concepts and applications of FL in healthcare and presents a novel approach for enhancing privacy preservation in Federated Learning.

Keywords. Artificial Intelligence, Federated Learning, decentralization, privacy

1. Introduction

The COVID-19 pandemic has stressed the need to improve our healthcare systems throughout the world. Many innovations in Artificial Intelligence (AI), mainly in Machine Learning (ML) and Deep Learning (DL) are being extensively applied in various domains of healthcare like disease diagnostics, drug discovery, clinical data prediction, patient monitoring, genome sequencing etc. Federated Learning (FL) is the latest development in AI which is related to building ML/DL models in a distributed, collaborative manner among participating entities without sharing their own raw data [1] [2]. The entities can be organizations like hospitals, banks, research centers, industries etc. or smart devices like sensors, wearables, mobile phones, laptops, tablets etc. FL has gained a lot of attention nowadays as it preserves data privacy and can be used to build AI applications in healthcare, smart manufacturing, computer vision, autonomous driving etc. The application of Federated Learning in healthcare is more relevant and necessary as it deals with private sensitive data of individuals which needs to be protected from unauthorized exposures. Many real-world projects of FL models in smart healthcare have been implemented worldwide and are expected to grow in coming years [5]. Even though FL preserves data privacy by inhibiting movement of raw data, some sensitive information can still be leaked from the exposure of parameters that are exchanged among participating entities. Therefore, a simple yet robust privacy preserving approach

[1] K Narmadha, Corresponding author, Department of Computer Technology, MIT Campus, Anna University, Chennai – 600044, India; E-mail: narmk27@gmail.com

for FL is proposed in this paper. Our proposed approach uses differential privacy in which noise is added to the data to hide its originality.

Section 2 gives introduction on the concepts of Federated Learning. A discussion on leveraging FL in healthcare is described in section 3. The proposed algorithm is presented in section 4. Finally, the paper ends with a conclusion in section 5 highlighting the directions which need more study and research.

2. The rise of Federated Learning

There has been an unprecedented growth of Big Data in recent years and it has further exploded due to the COVID - 19 pandemic. This has led to the growing demand for AI particularly ML and DL, in which computers are trained to process this enormous amount of data to derive meaningful information from it. In traditional method of building ML/DL models, the entire data from all data generating entities will be transmitted to the service provider's or server's location which is usually in a cloud. The complete process of training and building the model will take place only at the server's place. This method has a few challenges. The data which is needed to build a ML model comes from various scattered and geographically distributed sources. Often, it is tough to integrate this fragmented data from isolated remote locations due to factors like limited network connectivity etc. Moreover, a major part of the data involves private sensitive data of individuals. So, transmitting the private data across geographical boundaries in a secured manner and without violating privacy rights of the users is quite difficult. Also, transmitting huge amount of data to a single central location from sources across the world will increase the cost and choke the network bandwidth. On the other hand, higher the amount of data that is fed to a ML model, the accuracy of the model will be better. Thus, to overcome the challenges of data migration and integration as well as maintaining privacy of sensitive data, an alternate method called Federated Learning has attained importance in recent years.

Federated Learning is a type of distributed ML where the model is being collaboratively trained and built by the data generating entities (clients) which are either smart devices or organizations along with servers who act as coordinating entities [1] [2]. In FL, each client's raw data is used locally and not exchanged or transmitted to a server. All participating clients train a local ML model with their own data for a particular number of rounds until a qualifying criterion is met. After each iteration, they transmit the updated ML model parameters to the server for aggregation. The aggregated model parameters are transmitted back to all clients from the server which are used in the next iteration by the clients. This iterative process gets repeated until the model converges or the desired accuracy is obtained. This training process which takes place with one server and *n* clients is explained in the following steps:

- **Step 1:** Server initializes a global model GM and distributes to *n* clients.
- **Step 2:** Each client (C_i) initializes its local model (LM_i) as GM in first iteration.
- **Step 3:** Each client (C_i) trains its local model (LM_i) with its own data and transmits the model parameters to server.
- **Step 4:** Server does a weighted average of the local model parameters of all *n* clients and updates GM [10].

$$GM = \sum_i^n (W_i LM_i)$$ where W_i - Weight of ith client based on its dataset size

- **Step 5:** Server sends the updated GM to all *n* clients and the local models are modified with updated GM.
- **Step 6:** Each client (C_i) calculates the accuracy of the local model (LM_i) and verify whether the model has converged.
- **Step 7**: If the clients obtain desired accuracy or the local models have converged, the training process is stopped else the process continues from step 3 as next iteration.

Federated Learning (FL) preserves user data's privacy by default as the raw data does not leave the data sources. The domains like healthcare, military etc. which deals with sensitive data can make use of FL to reap AI benefits. Computational power needed for training a model is also shared among the clients instead of relying entirely on a server. Also, the organizations who do not have sufficient training data to build a standalone ML model can leverage FL to build a joint ML model. Moreover, network bandwidth consumption is significantly reduced in FL as huge amount of raw data is not transmitted to the server.

3. FL in healthcare

Federated Learning in healthcare is getting significant attention because of the need to build privacy-preserving and more accurate ML/DL models. FL in healthcare can be used for tasks such as to improve the prediction of diseases at an early stage, give best available treatment to patients, fasten drug discovery process etc. [8]. A handful of articles on FL in healthcare have been published by many authors [5][6][7][8][9]..

Federated Learning in smart healthcare can be used in three different scenarios depending on the type of participating clients.

- **Scenario 1**: Clients are organizations like hospitals, medical research centers, pharmaceutical companies, government medical bodies, genome sequencing labs etc. This type of FL is known as cross-silo FL [1]. For instance, two or more hospitals can collaborate and build a ML model with the help of a government medical body who can act as a server.
- **Scenario 2:** Clients are smart healthcare devices like wearables, smart patient monitors, sensors etc. This type of FL where the clients are smart devices used by individuals is known as cross-device FL [1]. For example, health parameters from wearables of thousands of users can be used to jointly train a ML model without moving the data out from the devices [7].
- **Scenario 3:** In this case, the clients can be either a healthcare organization or a smart healthcare device. This can be called as hybrid FL, where a collaboration between different hospitals and many standalone individuals is made to train and build a ML model.

The data partitioning among the clients can be horizontal or vertical. Two or more diabetes hospitals at different places can collaborate to train a ML model to predict the early stage of diabetes. In this scenario, all the hospitals share the same feature space and different sample space, which is known as horizontal FL [1]. There can also be a collaboration among a diabetes hospital, an eye hospital and a cardiology clinic at a same place. In this case, all the hospitals share different feature space and same sample space, which denotes a vertical partitioning of training dataset and is known as vertical FL [1].

This model can be used to find the relationships between diabetes, eye disorders and cardiac diseases. Federated transfer learning can be applied if a hospital and a pharmaceutical company want to jointly build a ML model to assess the side effects of drugs. Here, both the feature space and sample space differ among the clients.

The main advantage of using FL in healthcare is improved privacy protection of patients' sensitive data. Also, smaller hospitals who do not have sufficient training data and computational power can benefit from a joint collaboration. Moreover, accuracy of the model will definitely improve when the model is being trained in a federated manner as training takes place with huge amount of data.

3.1 Privacy in FL

Federated Learning preserves privacy of user's raw data, however some private information can still be exposed through model parameters that are exchanged between server and clients [4]. It is possible to extract raw data through model inversion attacks. Moreover, curious or malicious server and clients can try to infer other entity's information by analyzing the model parameters. So, different privacy-preserving mechanisms need to be applied to further improve the privacy preservation in FL. The three most popular methods that are used for privacy preservation in FL are homomorphic encryption, secure multiparty computation and differential privacy [4]. Among the three approaches, differential privacy is widely used in real-time applications as it is scalable and involves less overhead compared to the other two.

Differential privacy is a mechanism in which a little amount of random noise is added to any data to perturb its value. The amount of noise to be added depends on the sensitivity of the data. More amount of noise will reduce the utility of data as it leads to higher perturbation. On the other hand, it also achieves higher privacy protection. So an optimal amount of noise needs to be added to balance the trade-off between data utility and privacy. The noise value(N) is calculated from the Gaussian distribution.

$$N = (c.s) / \varepsilon \text{ where } c^2 \geq 2 \log (1.25/\delta) \text{ for } \varepsilon \in (0,1) \quad (1)$$

In the above equation (1), ε denotes the privacy budget and helps to control the level of noise added [4]. δ denotes the probability by which the ε-differential privacy is violated. Sensitivity(s) is calculated based on the maximum difference obtained on the model parameters when the model is trained on two neighboring datasets.

4. Proposed Algorithm

Server initializes a global model GM, ε, δ and distributes to *n* clients. Each client (C_i) initializes its local model (LM_i) as GM in first iteration. Each client (C_i) trains its local model (LM_i) with its own data, calculates the noise(N_i) based on equation (1), adds it with LM_i to get NLM_i and transmits NLM_i to the server.

$$NLM_i = LM_i + N_i \quad (2)$$

Server does a weighted average of the noise added local model parameters of all *n* clients, updates GM [10] and sends back to all n clients.

$$GM = \sum_i^n (W_i NLM_i) \quad (3)$$

where W_i - Weight of ith client based on its dataset size

Each client(C_i) subtracts its added noise(N_i) from GM and updates its local model for next iteration. Since the local noise is subtracted before proceeding with the next

iteration, the utility of the model parameters increases. Hence, our proposed approach produces better model accuracy.

5. Conclusion

Federated Learning in healthcare will see a tremendous growth in coming years. Some of the areas which need more study are vertical FL, decentralized topology in FL and privacy-preserving FL. Vertical FL in healthcare needs more concrete research as it can be used to find complex relationships among various diseases. Also, more research needs to be done for decentralized FL, which eliminates the requirement for a trusted server and thus improving the privacy preservation in FL. Technologies other than blockchain need to be explored in this type of FL. Effective and efficient privacy preserving mechanisms need to be studied to implement them in future real-world projects of FL.

References

[1] Qiang Y, Yang L, Tianjian C, Yongxin T. Federated Machine Learning: Concept and Applications, ACM Trans. Intell. Syst. Technol., 2019, Jan; vol. 10, no. 2, pp. 12:1–12:19.
[2] Li T, Sahu AK, Talwalkar A, Smith V. Federated Learning: Challenges, Methods, and Future Directions, IEEE Signal Processing Magazine, 2020, May; Volume: 37 Issue: 3, Page(s): 50 - 60.
[3] Kairouz P, et al. Advances and Open Problems in Federated Learning, 2019; Available: https://arxiv.org/abs/1912.04977.
[4] Wei K, et al., Federated Learning With Differential Privacy: Algorithms and Performance Analysis, IEEE Transactions on Information Forensics and Security, 2020, April; Volume: 15, Page(s): 3454 – 3469.
[5] Nguyen DC, et al. Federated Learning for Smart Healthcare. ACM Comput. Surv. 2021;Vol. 1.
[6] Pfitzner B, et al. Federated Learning in a Medical Context: A Systematic Literature Review. ACM Transactions on Internet Technology. 2021 May; Vol. 21, No. 2, Article 50.
[7] Yuan B, Ge S, Xing W. A Federated Learning Framework for Healthcare IoT devices. 2020, May; arXiv:2005.05083v1 [cs.LG].
[8] Brisimi TS, et.al. Federated Learning of predictive models from federated Electronic health records. International journal of Medical Informatics. 2018. Jan.
[9] Xu J, et.al. Federated Learning for Healthcare Informatics. Journal of Healthcare Informatics Research. 2020. November.
[10] McMahan HB, et.al. Communication – Efficient Learning of Deep Networks from Decentralized Data. 2017, Feb; arXiv:1602.05629v3 [cs.LG].
[11] Yu Y, et.al. Blockchain and Federated Learning for Privacy-preserved Data Sharing in Industrial IoT. IEEE Transactions on Industrial Informatics. 2020. June.
[12] Li X, et al. On the Convergence of FedAvg on Non-iid Data. 2020, June; arXiv:1907.02189v4 [stat.ML].

Challenges of Trustable AI and Added-Value on Health
B. Séroussi et al. (Eds.)

doi:10.3233/SHTI220437

Introducing Lightweight IT - A Way to Build Flexibility for Healthcare Organizations?

Gro-Hilde SEVERINSEN[a,1] Line SILSAND[a], Kristian MALM NICOLAISEN[a], Rune PEDERSEN [a,b] and Gunnar ELLINGSEN [a,b]
[a] *Norwegian center for e-health research, Tromsø, Norway*
[b] *Arctic university North Norway, Tromsø*

Abstract. The paper addresses socio technical relations of implementing a lightweight IT app in Norway. The focus is on how such technology will influence the possibilities for an organization connected to a large-scale heavyweight IT infrastructure to provide more rapid changes in line with healthcare worker's needs. The research question is therefore: How can lightweight IT support rapid organizational changes? The empirical site is one of the first health trusts in Norway implementing lightweight technology integrated to their Electronic Health Record (EHR). The lightweight technology is a nursing app for registering early- warning score (NEWS) bedside the patients. The paper is based on a qualitative interpretive approach, and the results are discussed in line with information infrastructure theory.

Keywords. Lightweight IT, Heavyweight IT, organizational change

1. Introduction

Large heavyweight IT constellations are becoming increasingly complex and rigid, requiring extensive time and effort to facilitate change and to meet innovation needs. Hence, such solutions make it is difficult for organizations to continue their digital development and adapt their IT portfolio to emerging user needs [1].

Increasingly, lightweight IT applications, complimentary to the existing heavyweight IT, is entering the e-health field. These innovative and agile applications that supports immediate user needs and work processes, with simple applications like apps [2]. The introduction and use of lightweight IT represents a paradigmatic shift in IT infrastructures in the Norwegian healthcare [2]. Hence, there is a need for knowledge for how this technology will affect users, clinical practices, and regional IT collaborations.

To address this knowledge gap, the paper uses a socio-technical perspective to investigate the implementation of a lightweight IT application. Previous research has shown how technology and devices can support clinical workflow, reduce clinicians time spend on transferring information, and reduce error in clinical documentation [3,4] This paper has a broader focus and investigate how lightweight technology can be a means for more rapid organizational changes in line with healthcare professionals' needs and emerging requirements. Our research question is therefore: *How can lightweight IT support rapid organizational changes?*

[1] Corresponding Author, Gro-Hilde Severinsen; E-mail: gro-hilde.severinsen@ehealthresearch.no.

The empirical site is the first health trusts (hereby called HT1) in Norway implementing a nursing app for registering early- warning score (NEWS) at the patient's bedside [5].

We adopt a theoretical perspective from the information infrastructure theory [6,7], and the principle of bootstrapping in relation to designing for instant usefulness, and connections to the existing system (installed base) [7].

2. Method

This paper is positioned within a qualitative interpretive paradigm, concerned with evolving and improving the understanding of a lightweight technology influencing organizational changes [9]. This is a qualitative inductive study, and the empirical data was collected one year after the nursing app was implemented.

The data included 22 semi-structured interviews with actors involved in the implementation and use of the app to get a broader overview of the socio technical implications. Informants include 26 healthcare professionals, 6 managers at HT1, 10 representatives from regional IT, and the vendors of the app and the EHR. The data collection also encompassed a demonstration of the app, and meetings and workshops with HT1. The interview guide encompassed questions on technology, work process and organization to provide an overview of the socio-technical aspects of the implementation process. The interview data was transcribed verbatim and subsequently categorized and analysed using systematic text condensation inspired by Malterud [10].

3. Results

The nursing app was implemented in 2020, with the aim of decreasing the use of paper, exploring mobile solutions in clinical settings, and simplifying the work with vital parameters and NEWS score. NEWS is a scoring tool used to detect and respond to clinical deterioration in adult patients and is a key element of patient safety and improving patient outcomes [5]. The score includes respiratory rate, blood oxygen saturation, temperature, systolic blood pressure, heart rate and a consciousness score [5]. In addition to improving the users' workflow and patient safety, the app was implemented to pilot digital procedures for clinical practice using mobile phones for registration, intended to increase digital competencies.

3.1. From Paper-notes to Nursing App

HT1 had priorly implemented a procedure of registering NEWS for all patients both in somatic and psychiatric care. This implied that the nurses had to observe and measure the patient's vital signs at the bedside and write the results of the measures on a paper note. Back in the ward office, they transferred the vital signs into a paper-schema and calculated the NEWS for each patient. The procedure of registering NEWS was conducted every morning before the doctor's ward round. The doctors had to go to the ward and search for the paper to update the NEWS status for their patients. Calculating NEWS implied double and triple registration of vital signs and the registration frequency increased if the patient status worsened. Accordingly, it was a time-consuming process

for the nurses, with an extensive risk of errors due to the complex transmissions of the data. Hence, HT1 sought to improve the workflow and procedure of registering vital signs and calculating NEWS, in addition to their overall aim of removing paper forms and making the data more accessible for both nurses and doctors. This required an app that was sufficiently user friendly and intuitive to use it in a busy clinical setting. In addition, it was necessary that the app could be integrated with the existing IT infrastructure and automatically transfer data to the EHR system. In early 2020, HT1 came to an agreement with a vendor offering a mobile app for registration of NEWS that met these requirements. The vendor had already established a collaboration with the regional EHR vendor to enable data exchange between the systems.

The app was piloted and implemented in an internal medicine ward before it was made available to the rest of the HT1. The feedback from the pilot ward was overall positive. Representing a potential barrier, the wards needed to actively request the app installed from the IT department. However, through a snowball effect more and more wards got the app installed. By January 2022, 26 wards, including almost all medical, surgical, and psychiatric wards, have installed the app.

3.2. The Adoption of the Nursing App – from Manual to Mobile Registration

Already in 2012, HT1 had implemented mobile phones as the solution for telephony, including emergency alerts and bed alerts. To improve the registration procedure for NEWS, HT1 had been looking for a solution to test out using the same phones for clinical work at the bedside. Most of the nurses found the mobile app intuitive and easy to use. The shift from paper schema to registering NEWS directly in the app did not demand extensive e-learning or classroom lessons. The IT department distributed a quick guide of instructions for using the app. The nurses accessed it from the existing mobile phones to register the vital signs necessary for the app to automatically calculate the NEWS score. The vital signs and the score were then automatically transferred from the app to the EHR system, as well as to the electronic whiteboard at the ward offices. The majority of the nurses instantly found the app useful and time saving since they did not need to register on paper anymore. However, some needed assistant to start registration of NEWS electronically due to low digital competence, and some described it as a barrier to use a mobile phone in front of patients. They worried that the technology would take the focus away from observing the patient assessment and communication with the patient. However, they solved it by explaining the purpose of using a phone at patient's bedside, to ensure that patients knew they were not "surfing on Facebook."

3.3. The Nursing App – Contributing to Improving Workflow

Since the functionality in the app was aimed for a specific clinical purpose, the implementation did not change the nurses existing workflow. They still had to measure and observe the patients to register the vital signs. However, the app had an evolving influence on work practices. By using the app, the nurse reported a better overview of the patients in relation to when to NEWS which patients and the total number of NEWS scores to do. They no longer depended on finding the paper chart or asking a colleague to know when it was time to do the next NEWS. Hence, the app ensured that everybody always had the same information about the patients' NEWS. The doctors highlighted that the patients' NEWS was available in the EHR from their office before ward rounds. This allowed the doctors to discuss the patients' conditions on doctors' meetings without

having to go get the paper chart first. In the EHR, they now got an important overview of each patient's NEWS trend over time, not only the latest NEWS score. Several of the ward managers described the app as not necessarily time saving for the nurses in terms of minutes saved for each registration. Instead, they emphasized how a busy clinical setting required the nurses to always prioritize their tasks. This meant that it often was challenging to find time to transfer vital signs scribbled down on a note, to the paper schema in the ward office, immediately after the measurements were taken. Hence, NEWS could be significantly changed without anybody knowing it for a long time. Using the app made this transfer automatically and immediately, thus improv the overall workflow for the healthcare professionals in the ward. After using the app for a while, the nurses at different wards required new schemas to be digitalized and installed in the app, e.g., the risk assessment forms for decubitus ulcers, nutrition and fall tendency, the forms for submission and admission, but also forms for more specific observations and scoring schemas tailored to a specialized hospital ward. The app was quickly adapted to the nurses' work practice, however, implementing it demanded for a complex overall regional collaboration.

3.4. The Relation Between the App and the Existing IT Infrastructure

Within the health region, HT1 was one of four health trusts participating in a regional IT collaboration. Hence, all implementations of clinical systems were conducted in a regional program where all health trusts had to agree upon the solution. The result was a complex and cumbersome processes, requiring years of negotiation and coordination before an app was implemented. However, the regional IT portfolio did not include all possible solutions the hospitals needed. The introduction of lightweight IT was therefore a step towards more flexibility for the health trusts, in terms of adding smaller solutions to the IT infrastructure without damaging the regional collaboration. HT1 had a mature IT organization and they had worked for years improving their IT portfolio and the work processes for their users. When HT1 found the app vendor for NEWS, they presented the solution at a regional level for approval.

In addition, the app vendor and the EHR vendor established a close collaboration to make the implementation, integration, and data exchange as smooth as possible. The app was integrated with the existing EHR system, and automatically transferred the NEWS data and generated a document in the EHR. The NEWS document was continuously updated with new registrations of vital signs. There was also a close collaboration between the vendors, the health trust, and the regional IT department on testing out and improve the mobile infrastructure to register clinical data and designing integrations to make the actual data exchange possible. All the actors involved had an interest in testing out mobile documentation for the healthcare domain which made this an efficient implementation process.

4. Concluding discussion

Traditionally, IT implementations in healthcare has consisted of introducing large-scale systems with a wide range of functionality for a heterogenous user group. Implementing complex solutions intended for addressing numerous user needs has proven cumbersome and demanding, challenging the realization of expected effects and benefits of the technology. Consequently, healthcare professionals have developed skepticism towards

new technological solutions. Implementing a mobile app designed for one specific purpose is a new way of introducing technology to support specific needs in clinical practice, without making large-scale changes to the existing II and the installed base [7].

The implementation of an app for documenting NEWS had a rapid and positive effect on the organization at different levels. First, in accordance with the bootstrapping principle [7], the app provided the users an instant value of improving the workflow without the need for demanding and comprehensive reorganization of work practices. In addition, the app generated changes in other parts of the organization, both related to the doctors' workflow and the overall overview at the wards. Even if the implementation of the app was done only to improve the NEWS procedure, it also generated new needs among the nurses. Second, since HT1 was part of a regional II, the work they did with testing and implementing the mobile app positively influenced the other health trust in the region as well. The installed base (the existing EHR) was used as the basis for how to integrate the app in relation to how information would be exchanged and distributed. This means that when another health trust wants to implement the app, it demands for minimum effort to their organization since all integrations to the EHR and collaborations between the necessary actors already are in place. This makes it possible for a health trust to complement their IT portfolio without large tender processes, complex implementations, and the risk of compromising the regional collaboration.

Although challenges in relation to how to procure and govern this new type of IT solutions for healthcare remains to be addressed, these solutions represent a promising step towards further digital development in healthcare. The fast adoption and the snowballing effect described in the present study, illustrates that lightweight IT has a potential to facilitate rapid organizational changes at different levels of healthcare. This includes, influencing several user groups and work processes both within the HT1 and in the overall regional infrastructure.

References

[1] Roland LK, Sanner TA, Aanestad M. Flexibility in EHR ecosystems: five integration strategies and their trade-offs. BIBsys. Open Journal Systems ISSN 1894-7719. 2017.
[2] Bygstad B. The Coming of Lightweight IT. ECIS 2015 ISBN 978-3-00-050284-2 http://aisel.aisnet.org/ecis2015_cr/22. 2015. 2015.
[3] Dall'Ora C, et al. How long do nursing staff take to measure and record patients' vital signs observations in hospital? A time-and-motion study. International journal of nursing studies 2021 118.
[4] Yange TS, and Onyekware U.O. An Implementation of a Point of Care Information System for Nurses. International Journal of Computer Applications. 2022 975: 8887
[5] Physicians, R.c.o. National Early Warning Score (NEWS) 2. Available from: https://www.rcplondon.ac.uk/projects/outputs/national-early-warning-score-news-2. 2022.
[6] Aanestad M, Jensen TB. Building nation-wide information infrastructures in healthcare through modular implementation strategies. The Journal of Strategic Information Systems. 2011 20(2): 161–176.
[7] Hanseth O, and Lyytinen K. Design Theory for Dynamic Complexity in Information Infrastructures: The Case of Building Internet. Journal of Information Technology. 2010 25 (1): 1–19.
[8] Walsham G. Interpretiv case studies in IS research: nature and methode. European Journal of information systems. 1995 4:74-81.
[9] Malterud K. Systematic text condensation: a strategy for qualitative analysis. Scandinavian journal of public health. 2012 40(8): 795-805.

Challenges of Trustable AI and Added-Value on Health
B. Séroussi et al. (Eds.)

doi:10.3233/SHTI220438

Suitability of Reusing Pre-Doctoral Student Activity Data from an Educational Information System for Quality Measures of Caries Risk

Benoit BALLESTER [ac,1], Estelle CASAZZA [a], Frédéric BUKIET [ab], Jean-Charles DUFOUR [cd]
[a] *Pôle d'Odontologie, Assistance Publique des Hôpitaux de Marseille, France*
[b] *Aix Marseille Univ, CNRS, ISM, Inst Movement Sci, Marseille, France*
[c] *Aix Marseille Univ, INSERM, IRD, SESSTIM, Sciences Economiques & Sociales de la Santé & Traitement de l'Information Médicale, ISSPAM, Marseille, France.*
[d] *APHM, Hôpital de la Timone, Service Biostatistique et Technologies de l'Information et de la Communication, Marseille, France*

Abstract. Dental caries management requires individualized follow-up and prophylaxis based on patients' caries risk (PCR). In large pre-doctoral clinics, the use of institutional quality measures (QMs) is essential to control the quality of patient follow-up and to evaluate the need for improvement measures. The aim of this retrospective study was to evaluate the suitability of reusing student activity data for the development of QMs of caries risk. Two approaches for predicting PCR using student activity data were evaluated and compared. The first approach used the procedure codes recommended by the Dental Quality Alliance and the second used these same codes along with three educational codes. The sensitivity, specificity, overall accuracy of the two approaches were evaluated. A Receiver Operating Characteristic (ROC) curve analysis was carried out, and the areas under the ROC curve of the two approaches were compared using Delong's test. A two-tailed P value ≤0.05 was considered statistically significant. While the two approaches were able to correctly predict PCR, the approach using both procedure and educational codes showed better predictive performance. The reuse of student activity data is an easy and robust method for the development of QMs of caries risk that can help improve monitoring and quality of patient care.

Keywords. Dental Informatics, Information System, Quality Measure, Dentistry, Education

1. Introduction

Dental caries prevention remains an important public health issue today. Still considered the most common chronic disease in the world [1], carious lesions can have significant adverse consequences for patients' health and functional status when left untreated [2]. However, since 2002, patients with dental caries have benefited from major

[1] Corresponding Author, Ballester Benoit, 27 Boulevard Jean Moulin, 13385 Marseille, France; E-mail: ballester.benoit.1@univ-amu.fr

developments in diagnosis [3], evaluation [4], and global management [5]. The latter involves individualized follow-up and prophylaxis based on patients' caries risk (PCR), which is currently defined as either high or low.

In pre-doctoral dental clinics, the global management of dental caries represents an organizational challenge because any malfunction can result in loss to follow-up and poor patient outcomes. In this context, the use of institutional quality measures (QM) appears essential to control the quality of patient follow-up and to evaluate the need for improvement measures. As early as 2011, the absence of QMs was identified as a barrier preventing oral health improvement, and in particular the reduction of oral health disparities [2].

In dentistry, the Dental Quality Alliance defined a set of standardized adult and pediatric QMs based on the Code on Dental Procedures and Nomenclature (CDT) to help assess oral healthcare access, process, and outcomes [6]. To date, most studies evaluating the use of QMs have employed administrative claims data [7,8]. Although the results are encouraging, access to and direct use of administrative data by dental practitioners is often difficult [9], notably in pre-doctoral clinic. Moreover, the development of QMs is impaired by the absence of diagnostic code usage in electronic dental records (EDRs).

At the Dental Department of Timone hospital (AP-HM - Assistance Publique - Hôpitaux de Marseille, France), as in other pre-doctoral clinics, care is mainly provided by dental undergraduate students under the close supervision of licensed dental practitioners [10] . Approximately 280 dental students handling about 12,000 patients each year. At the end of each patient visit, the procedures performed, and the skills acquired by the student are recorded in a structured way in an internally developed information system named ECHO. This information system then transmits the required billing information to the administrative department and allows the bi-annual evaluation of students.

The aim of this study was to evaluate the suitability of reusing pre-doctoral student activity data from the ECHO for the development of QMs of caries risk.

2. Materials and Methods

This retrospective study used student activity data stored in the ECHO information system of the Dental Department of Timone Hospital. This information system is registered in the CIL/AP-HM register under the number #2018-01 and contains, for each student, all the patients under care with the associated care activity and the validated skills. This study was registered in the RGPD/AP-HM register under the number #2021-61 and validated by the Ethics Committee of Aix-Marseille University (2021-06-03-11).

Patients were included in the study via ECHO based on the 2021 Dental Quality Alliance specifications of the QM "Percentage of children under age 21 years who have caries risk documented in the reporting year" (CRD-CH-A). Inclusion criteria were being aged between 1 and 21 years and undergoing continuous follow-up defined by at least two visits over 12 months with no interval greater than 31 days between them. A two-year inclusion period (from 2019/04/01 to 2021/03/31) was chosen.

Two approaches for predicting PCR status (high or low) using student activity data from ECHO were evaluated and compared. The first approach (ECHO Pro) employed the CDT procedure codes recommended by the Dental Quality Alliance as a proxy for PCR status. These codes were mapped to the procedure codes currently in use in France (CCAM - Classification Commune des Actes Médicaux). The second approach (ECHO

Pro +Edu) used the procedure codes above along with the following educational codes: (1) pulpotomy in patients aged over 16 years in a context of painful emergency; (2) root canal disinfection in a context of painful emergency; and (3) general anesthesia assistance in pediatric dentistry (usually performed for large-scale restorative procedures in high caries risk children). Patients for whom a procedure and/or educational code was reported in ECHO during the inclusion period were considered to be at high caries risk (no loopback approach was carried out).

The predictive performance of the two approaches was evaluated by determining their sensitivity, specificity, and overall accuracy. A Receiver Operating Characteristic (ROC) curve analysis was also carried out and the areas under the ROC curve (AUC) values of the two approaches compared using Delong's test [11]. To conduct these analyses, the PCR predicted using the two approaches was compared to the PCR reported in patients' EDRs, the latter being considered as the gold standard. The number of EDRs to be reviewed was estimated based on a prevalence of high PCR of 0.7 (as per preliminary data) with a type I error (α) of 0.05 and an accuracy of 90% [12,13]. Since we could not make an educated guess for sensitivity and specificity, we made the conservative choice of 50% [12]. The estimated number of EDRs to be reviewed was 321. The EDRs were randomly selected from the list of included patients until the number 321 was reached. As only free text notes were available in the EDRs, these were reviewed by two calibrated evaluators to identify the reported PCR.

Data were extracted from ECHO using PHP and MySQL scripts with CSV spreadsheet output. The PCR reported in EDRs were entered manually on a Microsoft Excel® spreadsheet. All analyses were performed with R for Windows® version 4.1.1. A two-tailed P value ≤0.05 was considered statistically significant.

3. Results

Over the inclusion period, 7,195 patients aged between 1 and 21 years visited the Dental Department of Timone Hospital. Of these, 2,261 met the inclusion criteria. A total of 384 randomly selected patient files had to be reviewed to obtain the 321 EDRs needed to evaluate the predictive performance of the two approaches (63 files were excluded due to incomplete or missing EDRs). The prevalence of patients at high caries risk was 0.67 (0.61-0.72). The predictive performance of the ECHO Pro and ECHO Pro+Edu approaches is presented in Table 1 with 95% confidence intervals.

Table 1. Predictive performance of the ECHO Pro and ECHO Pro+Edu approaches: True positive (TP); false positive (FP); false negative (FN); true negative (TN); prevalence (P); sensitivity (Se); specificity (Sp); overall accuracy (OA).

Approach	TP	FP	FN	TN	Se	Sp	OA
ECHO Pro	151	28	63	79	0.71 (0.64-0.72)	0.74 (0.64-0.82)	0.72 (0.66-0.77)
ECHO Pro+Edu	177	30	37	77	0.83 (0.77-0.88)	0.72 (0.62-0.80)	0.79 (0.74-0.83)

The overall accuracy was superior to the No Information Rate with a significant P-value of 0.03 for ECHO Pro and of 5.9e-07 for ECHO Pro+Edu. The ROC curve analysis (Figure 1) showed an AUC of 0.72 (0.67-0.77) for ECHO Pro and an AUC of 0.77 (0.72-

0.82) for ECHO Pro+Edu. The comparison of AUC values using Delong's test showed a significant difference in favor of ECHO Pro+Edu (P = 7.5e-05).

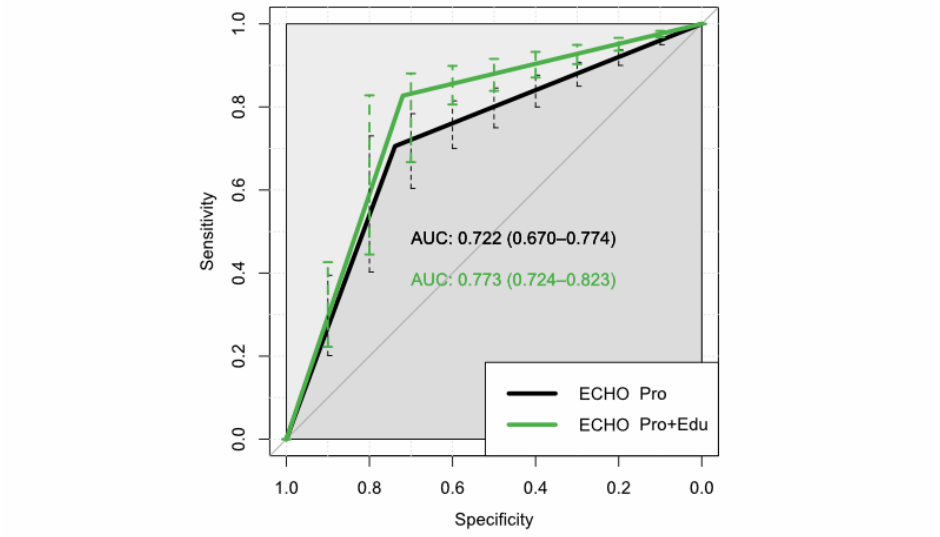


Figure 1. Receiver Operating Characteristics curve of the ECHO Pro and ECHO Pro+Edu approaches.

4. Discussion

In the absence of diagnostic code usage in EDRs in France, the standardized QMs of the Dental Quality Alliance can be used from procedure codes for the assessment of PCR. Although this required a mapping between CDT and CCAM procedure codes, the resulting test (ECHO Pro) were able to correctly predict PCR in more than 70% of cases, with a moderate P value for overall accuracy. However, the inclusion in the ECHO Pro+Edu approach of three educational codes linked to endodontic and restorative procedures yielded a higher sensitivity and resulted in better overall predictive performance. While the AUCs of the two approaches showed good predictive performance [14], a statistically significant difference between AUCs was observed in favor of ECHO Pro+Edu. In view of these findings, the approach using educational codes should be preferred.

A two-year inclusion period was chosen to limit the annual variability of results. The high prevalence of patients at high caries risk compared to the general population [15] may be explained by the fact that pre-doctoral clinics in France serve as referral centers that cater to disadvantaged populations. The percentage of patients with incomplete or missing EDRs was 16.4%. While this figure is consistent with published data [16], it does highlight the need to increase faculty and student awareness of the importance of properly filling patient records for medical and legal purposes. The fact that the ECHO is used for student validation may explain why it is more frequently completed and was able to identify and include patients for whom the EDR was incomplete or missing.

Our study has some limitations. The predictive performance of the two approaches may have been impaired by the lack of exact correspondence between the CDT and CCAM procedure codes. Unfortunately, we could not evaluate the effect on our results of mapping between these codes, as no study has evaluated the performance of the QM

"Percentage of children under age 21 years who have caries risk documented in the reporting year" (CRD-CH-A) to date—even though the latter is used as a basis for more complex QMs. Another limitation is that some of the endodontic and restorative codes can be reported in cases of trauma, which may have contributed to overestimating the prevalence of patients at high caries risk.

5. Conclusion

Our study evaluated two approaches for predicting PCR status using pre-doctoral student activity data. While the two approaches were able to correctly predict PCR, the approach using both procedure and educational codes showed better predictive performance. The reuse of activity data used for student evaluation has the advantage of being directly accessible by teaching staff and of being more often reported by students overall. Thus, reuse of student activity data provides an easy and robust method for the development of QMs of caries risk that can help improve monitoring and quality of patient care.

References

[1] GBD 2015 Disease and Injury Incidence and Prevalence Collaborators. Global, regional, and national incidence, prevalence, and years lived with disability for 310 diseases and injuries, 1990–2015: a systematic analysis for the Global Burden of Disease Study 2015. Lancet. 2016;388(10053):1545–602.
[2] Herndon JB, Tomar SL, Catalanotto FA, Rudner N, Huang I-C, Aravamudhan K et al. Measuring quality of dental care: Caries prevention services for children. J Am Dent Assoc. 2015;146(8):581–91.
[3] Gomez J. Detection and diagnosis of the early caries lesion. BMC Oral Health. 2015;15 Suppl 1:S3.
[4] Ismail AI, Sohn W, Tellez M, Amaya A, Sen A, Hasson H et al. The International Caries Detection and Assessment System (ICDAS): an integrated system for measuring dental caries. Community Dent Oral Epidemiol. 2007;35(3):170–8.
[5] Martignon S, Pitts NB, Goffin G, Mazevet M, Douglas GVA, Newton JT et al. CariesCare practice guide: consensus on evidence into practice. Br Dent J. 2019;227(5):353–62.
[6] Roberts MB. Dental Quality Alliance: measuring quality in dentistry. Tex Dent J. 2013;130(10):1006–9.
[7] Byrne MJ, Tickle M, Glenny A-M, Campbell S, Goodwin T, O'Malley L. A systematic review of quality measures used in primary care dentistry. Int Dent J. 2019;69(4):252–64.
[8] Righolt AJ, Sidorenkov G, Faggion CM, Listl S, Duijster D. Quality measures for dental care: A systematic review. Community Dent Oral Epidemiol. 2019;47(1):12–23.
[9] Riley GF. Administrative and claims records as sources of health care cost data. Med Care. 2009;47(7 Suppl 1):S51-55.
[10] Nalliah RP. Could dental school teaching clinics provide better care than regular private practices? J Investig Clin Dent. 2019;10(2):e12329.
[11] DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics. 1988;44(3):837–45.
[12] Buderer NM. Statistical methodology: I. Incorporating the prevalence of disease into the sample size calculation for sensitivity and specificity. Acad Emerg Med. 1996;3(9):895–900.
[13] Pourhoseingholi MA, Vahedi M, Rahimzadeh M. Sample size calculation in medical studies. Gastroenterol Hepatol Bed Bench. 2013;6(1):14–7.
[14] Šimundić AM. Measures of Diagnostic Accuracy: Basic Definitions. EJIFCC. 2009;19(4):203–11.
[15] Fleming E, Afful J. Prevalence of Total and Untreated Dental Caries Among Youth: United States, 2015-2016. NCHS Data Brief. 2018;(307):1–8.
[16] Tokede O, Ramoni RB, Patton M, Da Silva JD, Kalenderian E. Clinical documentation of dental care in an era of electronic health record use. J Evid Based Dent Pract. 2016;16(3):154–60.

Challenges of Trustable AI and Added-Value on Health
B. Séroussi et al. (Eds.)

doi:10.3233/SHTI220439

Establishing a Data Quality Baseline in the AKTIN Emergency Department Data Registry – A Secondary Use Perspective

Lucas TRIEFENBACH[a,1], Ronny OTTO[b], Jonas BIENZEISLER[a], Alexander KOMBEIZ[a], Saskia EHRENTREICH[b], Rainer RÖHRIG[a], and Raphael W. MAJEED [a] on behalf of the AKTIN Research Group
[a] *Institute of Medical Informatics, Medical Faculty, RWTH Aachen University, Aachen, Germany*
[b] *Department of Trauma Surgery, Otto von Guericke University, Magdeburg, Germany*

Abstract. Secondary use of clinical data is an increasing application that is affected by the data quality (DQ) of its source systems. Techniques such as audits and risk-based monitoring for controlling DQ often rely on source data verification (SDV). SDV requires access to data generating systems. We present an approach to a targeted SDV based on manual input and synthetic data that is applicable in low resource settings with restricted system access. We deployed the protocol in the DQ management of the AKTIN Emergency Department Data Registry. Our targeted approach has shown to be feasible to form a DQ baseline that can be used for different DQ monitoring processes such as the identification of different error sources.

Keywords. Data Accuracy, Software Validation, Electronic Health Records, Health Information Systems, Emergency Department

1. Introduction

Routine data from different medical electronic information systems such as electronic health records are increasingly re-used for secondary purposes such as patient care and research. While it offers the benefit of avoiding redundant data capture, it carries the risk of introducing a bias [1] e.g., from documentation practice and errors, making maintaining *data quality* (DQ) an ongoing issue. Root causes of DQ issues are variability of documentation practices, changes of processes and data integrity itself [2]. Secondary use of data thus requires an in-depth understanding of primary workflow processes [2].

Statistical analysis of data can identify systemic DQ issues [3] and DQ frameworks can help to implement feedback loops to increase DQ [4]. Thorough and continuous testing of productive electronic health records systems is a key to ensure adequate data quality. Test capabilities are generally available in such systems [5].

Audits and monitoring are used to ensure source data quality in clinical studies. Both methods employ source data verification (SDV), which requires adequate resources [6]. Even targeted source data verifications (TSDV) that partially cover the data sets increase

[1] Corresponding Author, Lucas Triefenbach, Institute of Medical Informatics, Faculty of Medicine, RWTH-Aachen, Pauwelsstr. 30, 52074 Aachen Germany, email: ltriefenbach@ukaachen.de

the workload for data management noticeably [7]. Increasingly, risk-based monitoring (RBM) methods are used to overcome limited resources [8,9]. RBM requires a baseline of data quality to continuously evaluate the quality of data, which can be found using SDV or TSDV [6,7]. If sensitive data is processed, legal restrictions may apply [10,11] and processes like SDV or RBM, which require access to data, are only possible using synthetic data [11]. The objective of this work was to develop a manual input-based protocol to determine a DQ baseline in settings with barriers to data access.

1.1. Background

The AKTIN *Emergency Department Data Registry* (AKTIN registry) provides a distributed and federated infrastructure for collecting and querying routine medical records for healthcare research [12], public surveillance and quality management [13]. The collected dataset is the *dataset emergency departmen*t defined by the *German Interdisciplinary Association for Intensive Care and Emergency Medicine [14]*. Currently, 37 emergency departments (EDs) are contributing de-identified data to the AKTIN registry, which is operated by the AKTIN research group. Participating EDs store records from the emergency department information system (EDIS) as a HL7 CDA document in a local data warehouse (DWH), from which they can be queried by the AKTIN research group after individual approval of the ED.

Ensuring data quality in the AKTIN registry requires a considerable amount of human resources, both on the side of the hospitals and on the side of the AKTIN research group. A main cause of technical errors is the mapping of variables during data export from the EDIS. The implementation, support and operation of the export interfaces is subject to contracts between the EDs and EDIS vendors. Interface testing options for the AKTIN research group are limited without direct system access, resulting in dependence on cooperating stakeholders to perform testing of interfaces or data storage. This lack of direct system access makes it difficult to distinguish DQ issue root causes e.g, whether issues occurred due to workflow mistakes, documentation practices or technical errors.

RBM is the approach used for monitoring DQ in the AKTIN registry. To determine a baseline of data quality, SDV is obliged. Preservation of data privacy opposes the use of real-patient data for SDV by third parties other than the ED. Thus, an approach by the AKTIN research group requires synthetic data. SDV corresponds to a black box testing method [15]. Since it is neither feasible to simulate various system environments with reasonable costs, nor to have direct access to productive systems, fully automated testing is not possible. Furthermore, the ED is a high stress environment with severe time constraints – therefore EDs have little resources for a full SDV. Our objective was to establish manual input based TSDV protocols for ensuring correct data transfer from EDIS to the AKTIN DWH, that are feasible in a low-resource setting.

2. Methods

2.1. Concept

Test execution is limited by the test input especially in the case of manual testing [15]. In the case of the AKTIN registry, the test inputs are executed by the ED staff providing

domain and system knowledge, which enables the execution of generic tests on specific systems. Thus, we minimize the ED staff workload in the test case modelling.

First, we examined the variables in the dataset emergency department to find a suitable subset of variables for full coverage. The data consist of 4 data types: free text, time stamps, numerical and finite discrete data fields. We did not consider free texts as unstructured data are not imported into the AKTIN DWH. We used equivalence class testing for timestamps and numerical values [15] i.e. we assumed that if representative values are correctly transmitted, all values of that variable are correctly transmitted. In the case of discrete data fields, equivalence classes were not applicable due to past issues with variable mapping. In this case, we considered all variable inputs.

Ordered by size, the 3 largest discrete variables were ICD-10-GM codes (over 70,000 codes) [16], CEDIS presenting complaints (CPC) (171 values) [14] and referral/discharge codes (12 values). Because full coverage of CPC and ICD-10-GM was not feasible manually, we did not include these variables for full coverage. All other variables were covered by 12 test cases resulting in a minimum test size of 12 test sets for execution. Additionally, we limited testing variables from 81 to 29 to reduce the time per test case. We based this prioritization on previous research queries and known issues from previous DQ monitoring. The covered variables included e.g., demographic data, triage system and score, vital signs, transfer codes, and associated time stamps.

2.2. Implementation

To reduce the workload of the medical staff, we reused known workflows for the test execution. Medical staff is familiar with registering patient consent using the AKTIN study infrastructure, which allows linking of de-identified patients with a study [12]. Hence, the test execution closely follows the same workflow as a study registration (c.f. fig. 1).

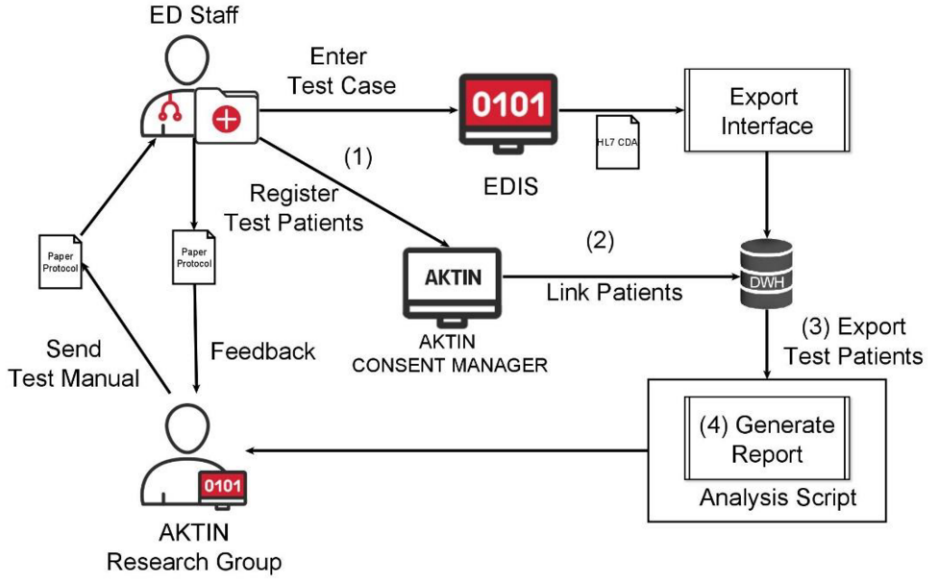


Figure 1. ED Staff receives test instructions from the AKTIN Research Group. They enter the patients in the ED Workplace System and register test patients in the AKTIN consent manager user interface (1) used to link

patients with studies (2). The Test patients are imported to the data warehouse (DWH) via the clinic export interface (3). Data are exported for analysis (4).

Staff received test data in form of synthetic patient records with instructions in a printable form. This form included a paper protocol for remarks. Staff entered the test patients in the EDIS and registered them as consenting patients in an AKTIN study created for the purpose of testing. After the input was completed, we queried test patients from the DWH and compared them against the expected results using an automated R script. We compiled a final report by adding information from the returned paper protocol manually. After that we removed the test patients from the productive DWH.

3. Results

So far, 17 clinics have participated and passed the SDV process. For these clinics, the process amounted to a workload of 1-2 person hours per medical staff and was met with positive feedback. In the process, we tested user interfaces and systems from 7 different vendors or in-house developments. In two systems from different vendors, linking of test patient with test protocols initially failed due to undocumented ID transformations. We resolved the problem in cooperation with IT-departments and system vendors. In all EDs, all variables were theoretically available except for data on ECG capture and laboratory orders. ECG and laboratory information are often located in subsystems and are not generated at EDIS level, thus being out of control of EDIS level test patients.

The DQ issues we observed were vendor-specific for the respective EDIS. Time stamps and numerical values were correctly transmitted in all systems. We observed issues with discrete variables. Some variables (I.e. isolation status) showed mapping errors, while variables related with internal processes such as mode of transportation, and transfer information showed mismatches between the AKTIN data set and the ED documentation practices/capabilities. Using the results as a DQ baseline, we can now differentiate technical errors from novel errors, i.e. workflow-related errors.

4. Discussion and Conclusion

We developed a protocol for generating a DQ baseline in the AKTIN registry, which is generally applicable for low-resource settings with limited data access. Reusing known workflows has been met with strong cooperation from involved parties and reduces the workload at input level. Solely the new process of linking test patient's IDs in the AKTIN study registration system proved to be more resource intensive than initially expected.

A mayor limitation is the design of the protocol. Manual input has to be provided by cooperating medical staff. While this restriction may make a TSDV practicable, it severely limits the scope of the tested variables. Additionally, limited staff resources for testing make our approach not suited for time critical applications. Specific testing staff might suit such scenarios, but at the same time disregard systematic errors by regular staff. While the use of synthetic test patients allows a high coverage per patient, test patients proved to have limitations especially regarding subsystems. Automatic input could address the input size issue but presents a larger organizational challenge.

The presented approach will be continuously used for generating a DQ baseline as presented e.g., for newly joining EDs in the AKTIN registry. In the future, other

approaches will have to address the pending issue of subsystem data and the covering larger parts of the data set.

We developed a protocol to evaluate baseline data quality in the AKTIN registry, demonstrating that it is feasible to use test-patients in productive systems such as EHR for TSDV. Our approach can distinguish between workflow related DQ issues and technical errors. While designed for the ED, our synthetic data-based approach is applicable anywhere where test patient input is possible.

5. References

[1] Martin-Sanchez FJ, Aguiar-Pulido V, Lopez-Campos GH, Peek N, Sacchi L. Secondary Use and Analysis of Big Data Collected for Patient Care. Yearb Med Inform. 2017;26(1):28–37.

[2] Ta CN, Weng C. Detecting Systemic Data Quality Issues in Electronic Health Records. Stud Health Technol Inform. 2019;264383–7. doi:10.3233/SHTI190248

[3] Kapsner LA, Kampf MO, Seuchter SA, Kamdje-Wabo G, Gradinger T, Ganslandt T, Mate S, Gruendner J, Kraska D, Prokosch H-U. Moving Towards an EHR Data Quality Framework: The MIRACUM Approach. Stud Health Technol Inform. 2019;267247–53. doi:10.3233/SHTI190834

[4] Ancker JS, Shih S, Singh MP, Snyder A, Edwards A, Kaushal R. Root causes underlying challenges to secondary use of data. AMIA Annu Symp Proc. 2011;201157–62.

[5] Wright A, Aaron S, Sittig DF. Testing electronic health records in the "production" environment: an essential step in the journey to a safe and effective health care system. J Am Med Inform Assoc. 2017;24(1):188–92. doi:10.1093/jamia/ocw039

[6] Houston L, Probst Y, Martin A. Assessing data quality and the variability of source data verification auditing methods in clinical research settings. J Biomed Inform. 2018;8325–32. doi:10.1016/j.jbi.2018.05.010

[7] Fougerou-Leurent C, Laviolle B, Tual C, Visseiche V, Veislinger A, Danjou H, Martin A, Turmel V, Renault A, Bellissant E. Impact of a targeted monitoring on data-quality and data-management workload of randomized controlled trials: A prospective comparative study. Br J Clin Pharmacol. 2019;85(12):2784–92. doi:10.1111/bcp.14108

[8] Barnes S, Katta N, Sanford N, Staigers T, Verish T. Technology Considerations to Enable the Risk-Based Monitoring Methodology. Ther Innov Regul Sci. 2014;48(5):536–45. doi:10.1177/2168479014546336

[9] Brosteanu O, Houben P, Ihrig K, Ohmann C, Paulus U, Pfistner B, Schwarz G, Strenge-Hesse A, Zettelmeyer U. Risk analysis and risk adapted on-site monitoring in noncommercial clinical trials. Clin Trials. 2009;6(6):585–96. doi:10.1177/1740774509347398 10.

[10] van Panhuis WG, Paul P, Emerson C, Grefenstette J, Wilder R, Herbst AJ, Heymann D, Burke DS. A systematic review of barriers to data sharing in public health. BMC Public Health. 2014;141144. doi:10.1186/1471-2458-14-1144

[11] James S, Harbron C, Branson J, Sundler M. Synthetic data use: exploring use cases to optimise data utility. Discov Artif Intell. 2021;1(1). doi:10.1007/s44163-021-00016-y

[12] Drynda S, Schindler W, Slagman A, Pollmanns J, Horenkamp-Sonntag D, Schirrmeister W, Otto R, Bienzeisler J, Greiner F, Drösler S, Lefering R, Hitzek J, Möckel M, Röhrig R, Swart E, Walcher F. Evaluation of outcome relevance of quality indicators in the emergency department (ENQuIRE): study protocol for a prospective multicentre cohort study. BMJ Open. 2020;10(9):e038776. doi:10.1136/bmjopen-2020-038776

[13] Otto R, Schirrmeister W, Majeed RW, Greiner F, Lucas B, Röhrig R, Walcher F, Brammen D. Implementation of Emergency Department Performance Benchmarking Using R and LaTeX. Stud Health Technol Inform. 2019;267238–46. doi:10.3233/SHTI190833

[14] Brammen D, Greiner F, Dormann H, Mach C, Wrede C, Ballaschk A, Stewart D, Walker S, Oesterling C, Kulla M. Lessons learned in applying the International Society for Pharmacoeconomics and Outcomes Research methodology to translating Canadian Emergency Department Information System Presenting Complaints List into German. Eur J Emerg Med. 2018;25(4):295–9. doi:10.1097/MEJ.0000000000000450

[15] Myers, Sandler C, Badgett T. The art of software testing. 3rd ed. Hoboken, N.J.: John Wiley & Sons; 2012. xi, 240 s. ;.

[16] BfArm. BfArM - FAQ/Ansprechspersonen [Internet] [cited 2022 Jan 21]. Available from: https://www.bfarm.de/DE/Kodiersysteme/Services/FAQ-Ansprechpersonen/_node.html German.

Challenges of Trustable AI and Added-Value on Health
B. Séroussi et al. (Eds.)

doi:10.3233/SHTI220440

Designing for Human Well-Being: A Case Study with Informal Caregivers of Individuals with Cancer

Awais AHMAD[a,1], Åsa CAJANDER[a], Birgitta JOHANSSON[b], Ylva TIBLOM EHRSSON[c], and Ulrica LANGEGÅRD[b]
[a] *Department of Information Technology, Uppsala University, Sweden*
[b] *Department of Immunology, Uppsala University, Sweden,*
[c] *Department of Surgical Sciences, Section of Otorhinolaryngology and Head & Neck Surgery, Uppsala University, Sweden*

Abstract. Informal Caregivers such as a spouse, other close relatives or friends of cancer patients can play an essential role in home-based treatment and care. However, the informal caregivers might not be prepared for this responsibility, and they might have several unmet requirements for taking care of patients in the home environment. The informal caregivers' physical, social and psychological health is also profoundly affected due to the health conditions of their relatives. We propose a User-centred Positive Design as a hybrid framework by merging the traditional User-cantered design and positive design frameworks to enhance the informal caregivers' subjective well-being. Our ongoing project (Carer-eSupport) will be used as a case study, and its main objective is to co-create and evaluate a web-based support system for informal caregivers of people with cancer. The proposed framework can be used for the design and development of health information systems with a special focus on users' wellbeing and positive emotions.

Keywords. Informal caregiving, Subjective well-being, Positive design, User-centered design, HCI, UCPD, Web-based support systems

1. Introduction

Cancer affects both the patients and their Informal Caregivers (ICs) [1]. ICs are usually the close relatives or friends of the patient who are responsible for caregiving without any payment. Caregiving to patients with cancer is a challenging and stressful task. The ICs usually are not well prepared for this responsibility and they lack the necessary information about receiving advice and help for home-based healthcare [1,2].

ICs' own physical and psychological health is also affected due to these life-changing circumstances. Many suffer from psychological problems such as depression, anxiety, and post-traumatic stress. ICs' burden of caring for cancer patients becomes intensive when the available resources and information are insufficient for caregiving in the home environment [3].

[1] Corresponding Author, Awais Ahmad; E-mail: awais.ahmad@it.uu.se.

In recent years, eHealth applications have emerged as efficient, cost-effective and easily accessible alternatives in healthcare [4]. Applications such as artificial intelligence-enabled and web-based support systems might be helpful for ICs' for caregiving individuals with cancer as well as to enhance their own subjective well-being. However, acceptance and adoption of these applications are still uncertain and further research is needed to investigate the proper design strategies that enable user participation and user well-being [5].

A user-centred design approach is widely used in human-computer interaction (HCI) research and beyond to design usable systems/tools that fulfil the users' needs[6]. The positive design framework (PDF) highlights the key elements of a person's subjective well-being [7]. In this paper, we propose the User-centred Positive Design (UCPD) by merging UCD and PDF. The merge between UCD and user's well-being has previously been explored by Gulliksen et. al. [8]. However, their work has focused on a professional work setting. The difference between those studies and this one is the context of informal caregiving and non-professional work setting.

Our proposed framework is expected to provide some clear guidelines for designing the eHealth systems that may support the users' subjective wellbeing and their positive emotions. We use our ongoing project (Carer-eSupport) as a case study; the project aimed to co-create and evaluate a web-based support system for informal caregivers of people with cancer.

2. Theoretical framework

2.1. User-centred Design

User-centered design (UCD) (also used as human-centered design) is an iterative process in which the users and their needs are considered in all steps of design and development. Norman and W. Draper stressed that addressing users' needs should be the main focus of the interface design of the system [9]. The primary purpose of UCD is to improve the usability and usefulness of the systems by using different techniques to involve the users in the design and development process of that system [10]. The circle in Figure 1 illustrates the four basic steps of the UCD process identified by the International standard organization (ISO).

2.2. Positive Design Framework

One of the primary reasons that motivate individuals to use a given technology is to improve their well-being by supporting their different needs[11]. The positive design approach describes how design can enhance the subjective well-being of the users. Desmet and Pohlmeyer suggested a three-dimensional framework for positive design (the triangle in Figure 1); the framework highlights three critical elements of subjective well-being: virtue, pleasure, and personal significance [7]. Desmet and Pohlmeyer argue that the suggested elements enable and stimulate human flourishing and positive emotions.

2.3. The Proposed User-centred Positive Design (UCPD)Framework

We propose a User-centred Positive Design (UCPD) framework, as shown in Figure 1, by combining the two frameworks PDF and UCD. UCD is important for a positive design, and the users' subjective well-being is a key element in it [7]. The users of a given design are the true evaluators of their subjective well-being. Desmet and Pohlmeyer stress that the system designer and the design process need to deliberately focus on human flourishing rather than just optimizing by finetuning the product/system and eliminating its flows [7]. In other words, a design should not only solve a given user's problem but should also make a long-lasting and positive, holistic impact on the user's well-being. Peters et. al also concluded in their study on designing for motivation, engagement and well-being that the impact of technology on human well-being can be and should be understood, designed and evaluated by focusing on the basic human psychological needs [12].

UCD gives us a systematic approach to fulfil the user's need, however, it does not provide the theoretical guidelines to understand the users and their needs. Due to an unclear understanding of human needs and different ways to apply UCD, the process leads to poor quality and usability of the system [8]. Therefore, we propose to use the UCD [10] as an implementation process and the PDF [7] as a theoretical framework. Figure 1 illustrates the proposed UCPD framework, in which we argued that the positive design elements should be focused in the UCD process, and the main purpose of UCD should be the subjective well-being of the users.

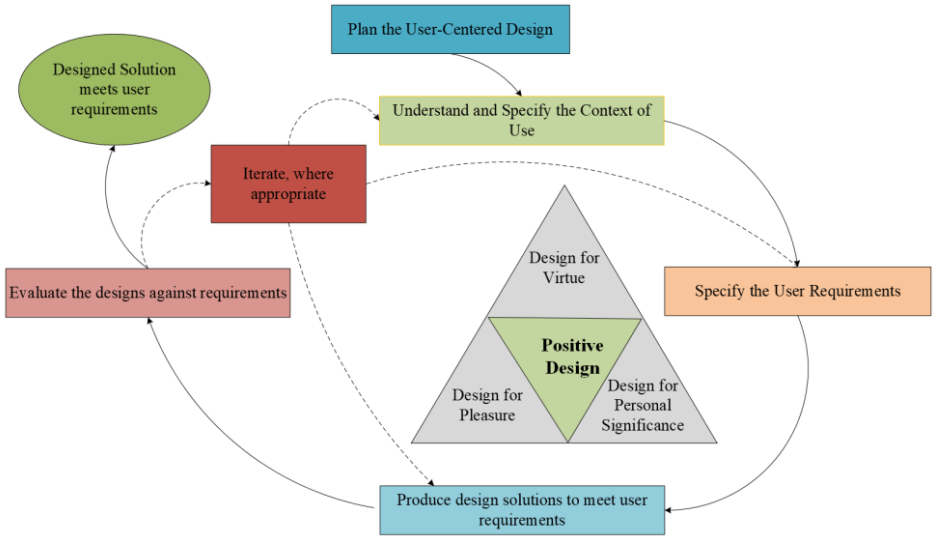


Figure 1. Proposed User-centered Positive Design Framework

3. The case study: Carer-eSupport project

We will use our ongoing project Carer-eSupport as a case study. We will test and adopt the proposed user-centered Positive Design (UCPD) framework in the project Carer-eSupport. The project aims to design, develop and evaluate a web-based support system

for the caregiving of cancer patients in the home environment. The intended system is also supposed to enhance the ICs' well-being and quality of life. We address the following research questions in this research project.

RQ1: What are the needs of informal caregivers of patients with cancer to support the patients in the home environment?

RQ2: How to design an Internet-based Support for Informal Caregivers to enhance their well-being and positive emotions?

RQ3: What are the critical factors influencing technology acceptance of Carer-eSupport?

3.1. Study Design

To ensure user participation, the UCD process will be used as research design in this project; and to enhance ICs' subject well-being, three elements of PDF: pleasure, virtue and personal significance will be used as theoretical guidance. Following UCD, interviews with the focus groups will be conducted to explore the users, their situation context and their perceived needs. Based on the interviews, an internet-based support system will be designed. The support system will be designed and developed in collaboration with U-CARE. The U-CARE provides an overall platform to deploy different eHealth services and it is used in several web-based interventions. The effects of the proposed support system on the ICs' well-being will be evaluated by a randomized controlled trial.

4. Discussion and Conclusion

In this paper, we shed light on the importance of users (ICs in the case study) well-being in the design and development of IT systems. UCD emphasize the user's involvement, and PDF advocates a holistic approach where human flourishing and subjective well-being should be the centre of any design approach. To support user's well-being and flourishment, we proposed the UCPD framework that merges PDF and UCD frameworks. In this study, we do not claim that the UCPD is a novel framework; after all, this is a merge between two already existing frameworks. However, the implementation of UCPD in our case study may lead to some substantial changes in the proposed framework, hence making it a novel framework in our future studies. The significance of this study is the context of the user as ICs.

Several studies attempted positive design strategies to support users' well-being [13–15]. To design eHealth solutions for the mental wellbeing of patients, Thieme et al. suggested a holistic approach where the key elements for mental wellbeing should be the user's psychological and social health, and positive emotions [14]. Cathy & Aidan used PDF to develop a design process called compassionate design; they suggest the design researchers should actively participate to develop the strategies that make a positive impact in the daily life of people who are suffering from different emotional and psychological issues [15].

To support users' well-being it is of utmost importance to involve the users in the design process of eHealth services [4]. However, the adoption of a systemic process to ensure the evolvement of actual users seems scarce in previous research. Our proposed framework gives clear guidelines to involve the users in the design and development of eHealth interventions with a special focus on users' wellbeing and positive emotions.

We are excited to adopt the proposed UCPD framework in Figure 1 for the Carer-eSupport project. Even though ICs physical and mental health and well-being is also deeply affected, the current interventions are focused on patients' caregiving and the ICs subjective well-being is seldom considered [3,16]. ICs need support for caregiving to the patients as well as for their own well-being and daily life activities. In the Carer-eSupport project, we strive to provide that support to ICs. Therefore, the aim of this project and the critical situation of ICs makes the project a perfect test case for the UCPD approach. We hope our proposed UCPD framework will become a potential solution for HCI and user experience design researchers to deliberately emphasize more on human flourishing in their design strategies rather than just solving a problem by fulling the user requirements.

References

1. Karampela M, Porat T, Mylonopoulou V. Needs of Head and Neck Cancer Patients and Stakeholders During Rehabilitation. In: Proceedings of the 13th EAI International Conference on Pervasive Computing Technologies for Healthcare. 2019. p. 415–21.
2. Adams E, Boulton M, Watson E. The information needs of partners and family members of cancer patients: a systematic literature review. Patient education and counseling. 2009;77(2):179–86.
3. Köhle N, Drossaert CH, Oosterik S, Schreurs KM, Hagedoorn M, van Uden-Kraan CF, Verdonck-de Leeuw IM, Bohlmeijer ET. Needs and preferences of partners of cancer patients regarding a web-based psychological intervention: a qualitative study. JMIR cancer. 2015;1(2):e4631.
4. Ahmad A, Mozelius P, Ahlin K. Factors Influencing Acceptance of Technology-enhanced Speech and Language Relearning for Stroke Survivors–A Systematic Review. In: 7th International Conference on e-Society, e-Learning and e-Technologies, Portsmouth, United Kingdom, June 10-12, 2021. 2021.
5. Guisado-Fernández E, Giunti G, Mackey LM, Blake C, Caulfield BM. Factors influencing the adoption of smart health technologies for people with dementia and their informal caregivers: scoping review and design framework. JMIR aging. 2019;2(1):e12192.
6. Campese C, Amaral DC, Mascarenhas J. Restating the meaning Of UCD And HCD for a new world of design theories. Interacting with Computers. 2020;32(1):33–51.
7. Desmet PM, Pohlmeyer AE. Positive design: An introduction to design for subjective well-being. International journal of design. 2013;7(3).
8. Gulliksen J, Göransson B, Boivie I, Blomkvist S, Persson J, Cajander Å. Key principles for user-centred systems design. Behaviour and Information Technology. 2003;22(6):397–409.
9. Norman DA. Cognitive engineering. User centered system design. 1986;31:61.
10. Iso DIS. 9241–210: 2010: ergonomics of human-system interaction—part 210: human-centred design for interactive systems (formerly known as 13407). Switzerland: International Standards Organization. 2010;
11. Zhang P. Toward a Positive Design Theory : Principles for Designing Motivating Information and Communication Technology. In 2008. p. 45–74.
12. Peters D, Calvo RA, Ryan RM. Designing for motivation, engagement and wellbeing in digital experience. Frontiers in psychology. 2018;9:797.
13. Vermeeren AP, Van Beusekom J, Rozendaal MC, Giaccardi E. Design for complex persuasive experiences: helping parents of hospitalized children take care of themselves. In: Proceedings of the 2014 conference on Designing interactive systems. 2014. p. 335–44.
14. Thieme A, Wallace J, Meyer TD, Olivier P. Designing for mental wellbeing: towards a more holistic approach in the treatment and prevention of mental illness. In: Proceedings of the 2015 British HCI Conference. 2015. p. 1–10.
15. Treadaway C, Taylor A, Fennell J. Compassionate design for dementia care. International Journal of Design Creativity and Innovation. 2019;7(3):144–57.
16. Heins M, Schellevis F, Rijken M, Donker G, van der Hoek L, Korevaar J. Partners of cancer patients consult their GPs significantly more often with both somatic and psychosocial problems. Scandinavian journal of primary health care. 2013;31(4):203–8.

Challenges of Trustable AI and Added-Value on Health
B. Séroussi et al. (Eds.)

doi:10.3233/SHTI220441

Multi-Dimensional Laboratory Test Score as a Proxy for Health

Bar H. EZRA [a], Shreyas HAVALDAR [b], Benjamin GLICKSBERG [c], and Nadav RAPPOPORT [a,1]

[a] *Department of Software and Information Systems Engineering, Ben-Gurion University of the Negev, Israel*

[b] *Department of Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai, New York, New York*

[c] *Hasso Plattner Institute for Digital Health, Icahn School of Medicine at Mount Sinai, New York, New York*

Abstract. The standard of care for a physician to review laboratory tests results is to weigh each individual laboratory test result and compare it to against a standard reference range. Such a method of scanning can lead to missing high-level information. Different methods have tried to overcome a part of the problem by creating new types of reference values. This research proposes looking at test scores in a higher dimension space. And using machine learning approach, determine whether a subject has abnormal tests result that, according to current practice, would be defined as valid – and thus indicating a possible disease or illness. To determine health status, we look both at a disease-specific level and disease-independent level, while looking at several different outcomes.

Keywords. Laboratory Tests, UK Biobank, Machine Learning, Electronic Health Records

1. Introduction

An essential part of the clinical decision-making process depends on interpreting different laboratory test scores relative to reference values. This process is usually performed by scanning the results and looking for abnormal values. Reference ranges have sev- eral limitations [1]. One of the core limitations is that the scan for abnormalities is done marginally, where every test is considered independently of the results of the other laboratory tests. Such practice can miss information in higher dimensions where a result of a laboratory test is assessed given the results of other laboratory tests. Another significant issue is that reference values are determined without a specific outcome in consideration, thus potentially creating a false display of abnormal values. Additionally, reference ranges are not always built on representative populations. Existing approaches have tried to cope with the reference values problem by developing different types of multivari- ate reference ranges [2,3]. Previous approaches used laboratory tests from hospitalized patients [4] or relied on longitudinal data [5], or tried to predict long-term effect [6].

[1] Corresponding Author: Nadav Rappoport; E-mail: nadavrap@bgu.ac.il.

We describe two approaches to cope with the limitations described above. The first is a disease-specific method, where the 'normality' of a set of laboratory tests for a given diagnosis is compared between affected and unaffected subjects. The second is a disease-independent method, using Machine Learning methods to detect abnormalities at a higher dimension point-of-view. The goal is to establish a method for alerting the physician that a patient is at a higher risk of having an illness or a potential risk and should be considered for further evaluation.

2. Materials and Methods

2.1. Datasets

Two datasets were used during this research – UK BioBank (UKBB) and data courtesy of the Mount Sinai Data Warehouse (MSDW).

The primary dataset used is the UKBBB which contains data of about 500,000 adult participants from all over the UK. This dataset consists of different types of features, such as demographic features, laboratory test values, previous and current illnesses as well as lifestyle, death information, and more [7]. This dataset was used as a baseline dataset. A total of 33 features including demographic data and laboratory test results were used as features, while disease diagnoses and admission data were used as outcomes of interest. There are known differences between sex in laboratory test results as well as in diseases' prevalence. Therefore, for the Disease-Specific model we split the analysis by sex and describe here females-based ansyslsis (267,746 participants). While for the Disease-Independent model, data from all participants were used. Participants with missingness were excluded, leaving 105,538 participants for the Disease-Specific model and 409,896 for the Disease-Independent model. Data was standardized by reducing the mean and scaling to unit variance by dividing by the standard deviation.

The secondary dataset that was used is the MSDW which aggregates clinical data from five hospitals within the Mount Sinai Health System in New York City. The wellness visits of the Mount Sinai Data Warehouse consists of about 100,000 visits for a wellness check-up. It contains demographic features, laboratory test values, and admission data, where the latter was used as the outcome of interest.

Laboratory tests with over 30% missingness were excluded. We were left with 31 different laboratory test types. Patients with missing data were excluded.

2.2. Disease-Specific Model

In this approach, we looked at a specific illness (e.g., Fatty Liver) and estimated the 'typical combination' of a set of laboratory tests. The objective was to test how well a multi-laboratory test-based score discriminates between ill and healthy patients. Subjects were split into two disjoint groups – ill and controls and were compared based on their scores. The score of subject i was defined as the Euclidian distance of the vector representing i's laboratory test results from the population's average. We hypothesis that this score is associated with illness.

$$Score(i) = d(\text{labs}_i, (labs)) \quad (1)$$

2.3. Disease-Independent Models

In this approach, we wanted to identify a general abnormality in the laboratory test scores by looking at the scores in a higher level of wellness. We trained Random Forest models using the following features age, sex, laboratory test results, and laboratory tests scores as defined above. We didn't use reference ranges to mark abnormal values to stay unbiased. The outcomes of interest represent general health measures as diagnoses and hospitalization rates. Data was split into 70% for models training and 30% for testing. Model's hyperparameter tuning was performed using cross-validation for each model to ensure the best results. Random Forest model was selected as it can capture non-linear effects and is robust to missing data.

We used two types of outcomes of interest: Number of diagnoses and number of admissions post laboratory test. The diagnoses-based outcome was defined according to the number of ICD10-CM codes assigned to each subject. Three binary outcomes were determined according to the number of diagnoses a subject had: at least 1, 5, or 10. A separate Random Forest model was trained for each of these outcomes. Admission-based outcomes were defined within two timeframes: whether readmission was recorded within 30 or 365 days. Due to high rates of imbalance of readmission outcomes, we trained a Balanced Random Forest model.

3. Results

3.1. Disease-Specific Score

We computed the distance of the representing vector of laboratory test results from the average vector for each subject. This distance was used as the subject's score. This analysis was performed for female subjects only. We hypothesized that our defined score is associated with the prevalence of specific diseases. The hypothesis was tested for six different diagnoses: Fatty Liver (ICD10-CM K760), Anemia (ICD10-CM D50*), Neutropenia (ICD10-CM D70*), Parkinson, Asthma, and Alzheimer.

Most subjects (266,511 out of 267,746) had at least one laboratory test outside the reference range. We found the cases and controls for all six diseases have a statistically significant difference in the distribution of the scores based on the Kolmogorov–Smirnov test. All p-values < 0.001 (Figure 1). Moreover, we found clear discrimination between cases and controls based on ROC AUC (Table 1). Moreover, to compare the current practice of marginal tests, we compared the discrimination of our score to the discrimination of other single laboratory tests using ROC AUC. We found that our score outperformed any other single laboratory test discrimination for Fatty Liver disease, Anemia, Neutropenia, and Asthma, but not for Parkinson's and Alzheimer's disease (Table 1). It should also be noted that the mean scores indicate that for Parkinson's and Alzheimer's disease, it is only slightly better than random classification.

3.2. Disease-Independent Score

In this method, we used the UKBB data as discovery data for the diagnoses prevalence model and for readmission rates models. The MSWD data was used for readmission models' validation, as this is the only outcomes information available in this dataset.

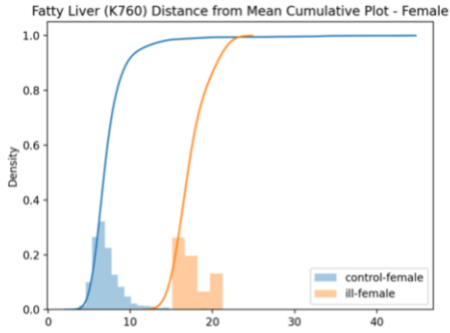


Figure 1. Distribution of the disease-specific scores among cases and controls for Fatty Liver.

Table 1. Comparison between top 5 Laboratory tests' ROC AUC scores and the Disease-Specific test ROC AUC score. GGT: Gamma Glutamyl Transferase, HGB: Hemoglobin, LYM: Lymphocyte, CREAT: Creatinine, EOS: Eosinophill, Glu: Glucose, MCH: Mean Corpuscular Hemoglobin, WBC: White Blood Cell, NEU: Neutrophill, HDL: High Density Lipoprotein, RBC: Red Blood Cell, Alb: Albumin, ALK: Alkaline, HCT: Hematocrit, MONO: Monocyte, LDL: Low Density Lipoprotein, MCV: Mean Corpuscular Volume, PLT: Platelet

Diagnoses	Top 1	Top 2	Top 3	Top 4	Top 5	DS-Score
Fatty Liver	GGT	Glu	HDL	ALK	HGB	**0.700**
	0.618	0.579	0.577	0.568	0.560	
Anemia	HGB	MCH	RBC	HCT %	MCV	**0.650**
	0.617	0.607	0.605	0.601	0.598	
Neutropenia	LYM %	WBC	LYM	NEU	PLT	**0.581**
	0.563	0.559	0.552	0.551	0.542	
Parkinson	CREAT	EOS %	LYM %	LYM	RBC	**0.525**
	0.541	0.533	0.529	0.523	0.515	
Asthma	EOS %	NEU	WBC	MONO	MCH	**0.551**
	0.542	0.516	0.512	0.512	0.511	
Alzheimer	CREAT	Cholesterol	Alb	LDL	LYM	**0.545**
	0.546	0.532	0.530	0.526	0.525	

Table 2. Performance of the three classification models. The prevalence and results are based on the 30% test set.

Observations	Prevalence	Accuracy	ROC AUC
At least 1 Logged Observations	74.3%	0.741	0.64
At least 5 Logged Observations	45.7%	0.633	0.68
At least 10 Logged Observations	26.0%	0.757	0.71

We created three different models for three binary outcomes based on a minimal number of diagnoses, 1, 5, or 10. As can be seen (Table 2), when looking at the ROC AUC score, increasing the threshold for cases definition increases the model's discrimination. We created two different models for each dataset for two different time horizons: readmission within 30 days and readmission within a year. The results from both datasets show that the model performs better when a longer time horizon is used (Table 3).

Table 3. ROC AUC results of the readmission rates model based on the 30% test set.

Timeframe	UKBB	MSDW
30 Days	0.64	0.68
365 Days	0.66	0.72

4. Discussion

In this study, we showed that using the shared information of common laboratory test results is advantageous over marginal information carried by individual laboratory tests. The advantage of a multi-dimensional score includes increased risk for specific diseases like Fatty liver diseases and for general health status as measured by the number of diagnoses and risk for readmission. Although we showed an association, it does not imply any causation. Meaning, we cannot conclude that shifting specific laboratory test results toward the mean or into the reference range will improve the health status. Using the number of diagnoses as an outcome of interest is not ideal, as the diagnoses come from the subject's history and are not limited to diseases diagnosed post laboratory test was taken. This limitation does not apply for readmission, where we limited to outcomes occurring post laboratory test taken. There exists other laboratory tests-based prediction models which are limited to hospital data or rely on longitudinal data [4,5] where the approached described in this work is based on routinely laboratory test results.

5. Conclusions

Considering the multi-dimensional distribution of common laboratory test results may be useful for alerting care providers of potential abnormalities. It can also be beneficial for predicting a specific illness and better or comparable to a single test assessment. The Disease-Independent score models can be used as an alert tool for physicians to indicate patients that may have been overlooked.

References

[1] Rappoport N, Paik H, Oskotsky B, Tor R, Ziv E, Zaitlen N, et al. Comparing ethnicity-specific reference intervals for clinical laboratory tests from EHR data. The journal of applied laboratory medicine. 2018;3(3):366-77.
[2] Boyd JC, Lacher DA. The multivariate reference range: an alternative interpretation of multi-test profiles. Clinical chemistry. 1982;28(2):259-65.
[3] Malka R, Brugnara C, Cialic R, Higgins JM. Non-parametric combined reference regions and prediction of clinical risk. Clinical chemistry. 2020;66(2):363-72.
[4] Froom P, Shimoni Z. Prediction of hospital mortality rates by admission laboratory tests. Clinical chemistry. 2006;52(2):325-8.
[5] Razavian N, Marcus J, Sontag D. Multi-task prediction of disease onsets from longitudinal laboratory tests. In: Machine learning for healthcare conference. PMLR; 2016. p. 73-100.
[6] Horne BD, May HT, Muhlestein JB, Ronnow BS, Lappe´ DL, Renlund DG, et al. Exceptional mor- tality prediction by risk scores from common laboratory tests. The American journal of medicine. 2009;122(6):550-8.
[7] Sudlow C, Gallacher J, Allen N, Beral V, Burton P, Danesh J, et al. UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. PLoS medicine. 2015;12(3):e1001779.

Challenges of Trustable AI and Added-Value on Health
B. Séroussi et al. (Eds.)

doi:10.3233/SHTI220442

Deviation of Physiological from Chronological Age Is Associated with Health

Lin PERETZ[a] and Nadav RAPPOPORT[a,1]
[a] *Software and Information Systems Engineering, Ben-Gurion University of the Negev, ISRAEL*

Abstract. Biological age may be of higher importance than chronological age, yet biological age is not trivial to estimate. This study presents a regression model to predict age using routine clinical tests like laboratory tests using the UK Biobank (UKBB) data. We run different machine learning regression models for this predictions task and compare their performance according to RMSE. The models were trained using data from 472,189 subjects aged 37–82 years old and 61 different laboratory tests results. Our chosen model was an XGboost model, which achieved an RMSE of 6.67 years. Subjects whose the model predicted to be younger than their actual age were found to be healthier as they had fewer diagnoses, fewer operations, and had a lower prevalence of specific diseases than age-matched controls. On the other hand, subjects predicted to be older than their chronological age had no significant differences in the number of diagnoses, number of operations, and specific diseases than age-matched controls.

Keywords. Biological age, Machine Learning, Laboratory tests, Chronological age, Electronic Health Records, BioBank

1. Introduction

As the world's aging population grows at an unprecedented rate, there is a clear need to learn more about the biological aging process and the determinants of healthy aging. To achieve this goal, researchers seek biological markers and other factors that can track biophysiological aging and, ideally, provide insight into the underlying mechanisms [1,2,3,4].

Biological age (BA), also called physiological age, measures how well or poorly a person's body functions. BA is correlated with calendar age (CA), also called chronological age, which is an objective measure of elapsed time since birth. The BA of a person can be higher or lower than their CA, since aging is not only a matter of time but is, in fact, a complex process with multiple causes. Studies have shown that young individuals of the same chronological age varied in their BA. These individuals showed cognitive decline and brain aging, self-reported worse health, and looked older [5].

There are several ways for determining BA, but none are definitive or truly accurate. Previous studies used a variety of ways to estimate the BA of a person. For example, studies used human physical activity as recorded by a wearable device [1], by cognitive variation independently of chronological age [6]. Other studies used molecular

[1] Corresponding Author, Nadav Rappoport, Department of Software and Information Systems Engineering, Ben-Gurion University of the Negev, ISRAEL, 1 Ben-Gurion Boulevard, Beer-Sheva, ISRAEL; E-mail: nadavrap@bgu.ac.il.

biomarkers like telomeres' length and others [1,2,3,4]. These approaches for estimating BA are limited as it requires the active collection of data specifically for this purpose and is costly.

This study presents a predictive model of chronological age applied to the UK Biobank (UKBB) data. The type of data used is often routinely collected laboratory test results. The advantage of using routinely collected laboratory results for estimating BA is that such a resource is very readily available, easy, and cheap to execute and does not need to be actively collected for our purpose.

Laboratory test results from the UKBB were used as features to train different machine learning regression models, and we chose the model with the lowest RMSE on a 25% held-out test set. This model has good performance in predicting the subjects' age, but the model makes significant errors in some cases. The cases where there is a large deviation between the predicted age and the actual age are the focus of this study. These cases were divided into two groups. Group P-Y (predicted younger) for which the model predicted to be younger than their actual age, and Group P-O (predicted older) for which the model predicted to be older than their actual age. On these deviation groups, we conducted further investigation determining their overall health vis-à-vis the rest of the population and vis-à-vis their age-matched groups.

In another study, Zichen et al. used vital signs and lab tests from an EMR to predict CA [7]. They have found specific diagnoses enriched in subjects with a high discrepancy between CA and BA.

Our study focused on adults population and demonstrated the use of BioBank data to assess the overall health, using different measures, of the population with a high discrepancy between CA and BA.

2. Methods

2.1 Data

The UK Biobank (UKBB) is an open-access research database and a large prospective study with about 500,000 adults (aged 40-69 at recruitment) participants recruited from 2006 to 2010. This database covers thousands of clinical and environmental variables such as demographic features (e.g., sex, age), laboratory test values, previous and current illnesses, lifestyle, imaging data, hospital discharges, and more [8]. We used data from 472,189 subjects consisting of 61 types of laboratory tests and the subject's gender and chronological age (CA). Missing values laboratory tests values were imputed using the median.

2.2 A predictive model for aging

To find the best model to predict the CA, we trained different machine learning regression models and compared their performance according to the root mean squared error (RMSE). The models were trained on a 75% random sample, and performances were evaluated using the 25% held out test set.

2.3 Identifying individuals with a significant discrepancy between predicted and chronological age

To identify subjects with significantly older or younger predicted age than their CA, we retrain our model on all the data without splitting it to train and test. Subjects for whom the model predicted a difference of more than two standard deviations from their CA, were considered as cases. Cases were divided into two groups: (i) cases with a prediction younger than their CA (P-Y); (ii) cases with a prediction older than their CA (P-O). We created two age-matching control cohorts. P-Y's control group contains subjects with CA of 67-82, matching CA range of group P-Y. P-O's control group contains subjects with CA in the range of 40-47, which is the CA of group P-O.

2.4 Outcome's comparisons

We used two sets of outcomes of interest, Overall health measures and disease-specific. We used several measures for estimating the overall health of each group. We used inpatient data available for a subset of the UKBB population. The data include admissions, diagnostic, and operation codes (UKBB data fields 41272, 41211, and 41270, respectively). Diagnoses were considered in two ways. In the first, we counted the number of ICD-10 diagnoses codes for each subject. In the second, we counted the number of groups of diagnoses codes based on Clinical Classifications Software (CCS) [9]. Subjects for whom clinical data were not extracted from electronic health record systems were excluded. After this step, we were left with 413,828 subjects, 177 in group P-O and 343 in group P-Y.

We compared the number of diagnoses, admissions, and operations of the two cases groups versus their age-matched groups. We tested the differences between the groups using the non-parametric Kolmogorov–Smirnov (KS) test to examine the hypothesis that there are statistically significant differences between the P-Y/P-O groups and their age-matched groups.

To perform the disease-specific analysis, we extracted diagnoses for a set of common diseases. Disease extraction was conducted in two different ways. First, we used the health-related outcomes category from the UKBB and extracted diseases with specific data filed. Second, we extract additional diseases using the ICD10 codes. We performed a chi-squared test to test the hypothesis that there are statistically significant differences between P-Y/P-O groups versus their age-matched groups and controlled for multiple testing using the Benjamini Hochberg method.

3. Results

We trained different machine learning regression models and computed their performance using RMSE. The Linear Regression model had the lowest performance (RMSE=7.214) compared to more complex models such as Random Forest (RMSE=6.803), XGboost (RMSE 6.271), and Deep Learning model multilayer perceptron (MLP) (RMSE=6.389). We chose to use the XGboost model as our age prediction model due to its lowest RMSE.

Hyper-parameters tuning was performed using a grid search cross-validation approach. The chosen parameters were learning rate of 0.03, maximum depth of a tree equal to 12, minimum sum of instance weight equal to 15, number of estimators equal to

500, and subsample equal to 0.75. The RMSE of this model on the test set was 6.271 years.

3.1 Outcome's comparisons

The two groups of interest were compared to the matching controls using inpatient data. We found that the P-Y group had significantly fewer diagnoses, group of diagnoses, number of operations, and number of admissions. On the other hand, the P-O group showed no significant differences compared to their age-matched group (Table 1).

Next, we compared differences in the prevalence of specific diseases between the groups, and we applied Benjamini Hochberg's correction of p-values for multiple tests [10]. First, we compared group P-Y with their age-matched and found that they had statistically significant lower prevalences of all 16 tested diseases (Table 2). Second, we compare group P-O with their age-matched and didn't see significant differences between the groups (Table 2).

Table 1. Outcomes comparison between the groups. Mean (SD) and the p-value of the KS test results. The group of diagnosis codes is based on CCS classification (see Methods).

	Mean (SD)				P-value	
	P-O	**40-47**	**P-Y**	**67-82**	**P-O vs. 40-47**	**P-Y vs. 67-82**
#Diagnoses	6.98 (7.6)	6.94 (8.6)	6.59 (5.8)	11.74 (12.2)	0.27	**1.26e-12**
#Grouped Diagnoses	5.69 (5.5)	6.27 (6.7)	5.66 (4.6)	9.28 (8.4)	0.42	**3.65e-12**
#Operations	6.51 (6.5)	6.39 (6.6)	6.80 (5.4)	9.76 (8.6)	0.99	**2.74e-07**
#Admissions	6.98 (7.6)	6.94 (8.6)	6.59 (5.8)	11.74 (12.2)	0.27	**1.26e-12**

Table 2. Prevalence of different diseases in 4 groups of subjects. Prevalence is given in percentages. The difference between the affected group and its matched controls is adjusted for multiple false discovery rate using the Benjamini-Hochberg method.

Disease	**Group P-Y (n=343)**	**Age-matched (n=64,550)**	**P-value**	**Group P-O (n=177)**	**Age-matched (n=61,610)**	**P-value**
Cancer	24.78%	26.85%	0.504	9.60%	8.66%	1
Asthma	9.03%	13.00%	0.064	19.20%	15.53%	1
COPD	2.04%	6.41%	0.003	3.38%	1.37%	0.784
Dementia	0.29%	1.55%	0.15	0.00%	0.06%	1
ESRD	0.00%	0.26%	0.707	0.56%	0.18%	1
Motor Neuron	0.00%	0.13%	1	0.00%	0.02%	1
Heart Attack	4.08%	7.88%	0.027	2.25%	1.36%	1
Parkinson	0.29%	0.94%	0.482	0.00%	0.08%	1
Stroke	1.45%	4.79%	0.016	1.12%	1.05%	1
Hypertension	32.65%	50.53%	8.76E-10	14.68%	14.36%	1
Anemia	1.45%	4.16%	0.036	2.25%	2.94%	1
Atherosclerosis	0.29%	0.81%	0.504	0.00%	0.09%	1
Chronic liver disease	0.00%	0.45%	0.504	0.00%	0.40%	1
Diabetes	1.45%	9.77%	1.89E-06	5.64%	3.40%	1
Heart disease	11.37%	27.01%	8.76E-10	4.51%	5.73%	1
Kidney disease	1.45%	7.17%	2.60E-04	2.25%	1.36%	1

4. Discussion

A machine learning regression model was shown to be valuable for predicting age based on laboratory test results. Our model can predict subjects' CA with an RMSE of 6.27 years. The cohort of subjects with estimated BA significantly lower CA has overall better health than the age-matched group. The overall health was approximated using the number of diagnoses, operations, and admissions. In addition, this population was found to have lower prevalences for all 16 tested diseases. On the other hand, we found that the cohort for which the BA was higher than their CA had no significant differences.

This study has a few limitations. First, the cohort is composed of only adults subjects, and its application to a broader range of ages is yet to be proven. In addition, the models are based on tens of different laboratory tests, which limits its broad use. Moreover, in this study, we showed that there is an association between the deviation of BA from CA and health. But this does not imply causation, nor any suggestion for intervention that can improve health.

5. Conclusion

We demonstrated the usefulness of a machine learning model for age prediction based on laboratory test results. We showed that subjects with a significantly lower predicted age than actual age were 'healthier'.

References

[1] Syed Ashiqur R, Adjeroh DA. "Deep Learning Using Convolutional LSTM Estimates Biological Age from Physical Activity." Scientific Reports 9, no. 1 (December 2019): 11425.
[2] Janet L, McGill NI, Lindsey LA, Green DK, Cooke HJ. "In Vivo Loss of Telomeric Repeats with Age in Humans." Mutation Research/DNAging 256, no. 1 (January 1991): 45–48.
[3] Bae CY, Kang YG, Kim S, Cho C, Kang HC, Yu BY, et al. "Development of models for predicting biological age (BA) with physical, biochemical, and hormonal parameters". Archives of Gerontology and Geriatrics, 47, 2, September–October 2008.
[4] Belsky DW, Caspi A, Houts R, Cohen HJ, Corcoran DL, Danese A, et al. "Quantification of Biological Aging in Young Adults." Proceedings of the National Academy of Sciences 112, no. 30 (July 28, 2015): E4104–10.
[5] Åke W, MacDonald SWS, de Frias CM, Nilsson LG, Dixon RA. "How Do Health and Biological Age Influence Chronological Age and Sex Differences in Cognitive Aging: Moderating, Mediating, or Both?" Psychology and Aging 21, no. 2 (2006): 318–32.
[6] Zichen W, Li L, Glicksberg BS, Israel A, Dudley JT, Ma'ayan A. "Predicting Age by Mining Electronic Medical Records with Deep Learning Characterizes Differences between Chronological and Physiological Age." Journal of Biomedical Informatics 76 (December 2017): 59–68.
[7] Sudlow C, Gallacher J, Allen N, Beral V, Burton P, Danesh J, et al. "UK Biobank: An Open Access Resource for Identifying the Causes of a Wide Range of Complex Diseases of Middle and Old Age." PLOS Medicine 12, no. 3 (March 31, 2015): e1001779.
[8] Tianqi C, Guestrin C. "XGBoost: A Scalable Tree Boosting System." Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, August 13, 2016, 785–94
[9] Salsabili M, Kiogou S, Adam TJ. "The Evaluation of Clinical Classifications Software Using the National Inpatient Sample Database," n.d., 10.

Challenges of Trustable AI and Added-Value on Health
B. Séroussi et al. (Eds.)

doi:10.3233/SHTI220443

Design Considerations for the Use of Patient-Generated Health Data in the Electronic Medical Records

Katherine BLONDON[a,b,1] and Frederic EHRLER[a]
[a] *University Hospitals of Geneva, Switzerland*
[b] *University of Geneva, Swtizerland*

Abstract. Patient-generated health data (PGHD) is of growing interest to physicians, particularly if they are integrated in the electronic medical record (EMR). Concerns about how to manage vast amounts of PGHD and potential liability issues have limited their use. Based on interviews with specialists, we present types of PGHD, workflow processes and needs. We then discuss consideration for how to manage PGHD with approaches for analyses to detect abnormal results, and present implications for alert systems and visualization requirements in multi-patient views.

Keywords. Patient-generated health data (PGHD), alerts, patient-reported outcome (PRO)

1. Introduction

With the widespread adoption of smartphones, tracking devices and connected devices, patients collect vast amounts of data about their health. Patient-generated health data (PGHD) include several types of data. Some require patient inputs such as clinical parameters (e.g., blood pressure or glucose measurements), or patient-reported outcomes (typically surveys or questionnaires). In fact, the definition of PGHD by the Office of the National Coordinator for Health Information Technology (ONC) [1] also includes health data collected from family members or other caregivers. Healthcare providers may encourage their patients to collect and share their health data to help manage a medical issue. They may send questionnaires before or after a medical visit or may want to help patients with their self-management.

Since PGHD can potentially generate vast amounts of data, providers worry about receiving too much data from their patients, and not have time to process these data, with concerns about subsequent liability for abnormal findings. Artificial intelligence can provide solutions to handle the increasing amount of data made available by PGHD, by detecting anomalies and sending alerts, for example.

[1] Corresponding Author, Katherine Blondon, Medical Directorate, University Hospitals of Geneva 1205 Geneva, Switzerland; E-mail: Katherine.blondon@hcuge.ch.

In a prior paper, we focused on liability and interoperability issues of PGHD [2] in this paper, we aim to explore design considerations to detect and manage abnormal results and to create an adapted alert system. We therefore interviewed healthcare professionals to identify the PGHD of interest and their usage to analyze how types of data will determine different models of interpretation to create adapted alerts.

2. Methods

Our understanding of PGHD is based on individual semi-structured interviews and focus group with healthcare professionals from several contexts of care. These encounters were conducted to understand clinicians' use and needs for PGHD, as well as concerns in implementing this kind of data in our EMR.

We asked clinicians about the type of data they wish to collect about their patients, how they intend to collect it, how they would like to see it in relation to the EMR, and their concerns about collecting PGHD. We conducted a thematic analysis of the collected transcripts.

We report our findings in the results and discuss the feasibility of the different solutions for a safe and reliable integration into the EMR. We reviewed the various approaches to managing PGHD, analyzing possible triggers for alerts (e.g., thresholds, pattern recognition, and trends with various methods of analyses). We conclude with design implications for visualization and alerts systems for PGHD.

3. Results

From June 2021 to December 2021, we encountered a total of 14 physicians, 2 nurses and one project manager from 12 specialties: oncology, cardiology, infectious diseases, nephrology, diabetes and endocrinology, psychiatry, family (ambulatory) medicine, emergency medicine, general surgery, neurosurgery, pain specialists (from anesthesiology) and telemedicine. Four individual semi-structured interviews and two focus group were conducted.

The thematic analysis identified the following topics: types of PGHD, visualization in the EMR, alerts and workflow.

3.1. Types of data and data collection

Four specialties currently already use patient questionnaires (paper or electronic surveys), two only use questionnaires for research, and five divisions are still considering how to implement patient-reported outcome (PRO) tools. None of the current electronic questionnaires are integrated with our EMR. Some divisions review patient-generated data on websites or import them from patients' devices (e.g., glucose measurements).

Five specialties anticipated 2 to 10 new patient questionnaires in the near future, which could target 40 patients in medical specialties to 3000 in pre-hospital consultation patients per month. All participants were interested in questionnaires for their patients, with different objectives: these ranged from preparing a medical visit,

following up after a visit or for research purposes. For example, the oncologists use a questionnaire to assess for side-effects of chemotherapies [3].

In addition to use questionnaires, four specialties were interested to collect (or plan to collect) PGHD from a variety of devices to adapt the care for their patients. Glucometers, continuous glucose devices, scales, pedometers, etc. were commonly considered sources of PGHD.

3.2. Workflow and management of PGHD

All participants agreed that management of PGHD would require a new workflow. An initial workflow is required to send the right questionnaire to the right patient at the right time. The clinicians needed to be able to "prescribe" the questionnaire or have an easy process to ensure that a given patient would receive the right questionnaire. Patients may need support for technical or clinical assistance with the questionnaire. After the patient submits her responses, there needs to be another process to ensure that the responses are seen, whether the results are in the EMR or not.

Finally, results of the questionnaire need to be analyzed to determine the appropriate action: urgent responses need immediate attention such as contacting the patient or alerting the patient's doctor. It may result in a new appointment, lab test, or imaging. The urgency of managing the responses will depend on the responses the patient has given. The whole questionnaire may be a clinical score, which needs to be interpreted as a single entity, other questionnaires may have key questions that may require an adapted response. An example of key question is a self-assessment of a suicidal risk, which may require an urgent response.

3.3. Visual design and dashboards

All participants agreed on the need for a summary view of patients and results. Medical assistants or other providers who work with PROs need to have an overview of sent and answered questionnaires to follow up on their patients in a multi-patient view. This is the initial dashboard requirement, which should also show the various alerts that we discuss below. Visual design needs to distinguish process alerts (no answer, for example) from content alerts, with varying levels of urgency.

When abnormal findings of low urgency need to be seen at the patient's next visit, the alert needs to be visible at the right time for the given user: our interviewed physicians were very concerned about alert fatigue and information overload, leading to missed information. Therefore we address considerations for designing alerts in the next section.

4. Discussion

4.1. Considerations for designing alerts

Some parameters have simple interpretations, whereas other need to integrate additional information for interpretation. When an abnormality is defined by a single question, or a final score, an alert system is easily designed, triggered by the abnormal result. The subsequent question is to define who needs to receive the alert. When we

asked each physician how to design the alerts and who to target, we received different responses, according to the degree of urgency. These are summarized in Table 1.

Table 1. Dispatch of alert and their degree of urgency

Degree of urgency	Example	Action	Alert receiver
Extremely urgent	Recurrent chest pain after myocardial infarction	Call for an ambulance	Medical assistant or nurse, Doctor
Urgent	Recurrent hypoglycemia at 6pm over 3 days, when driving home from work	Contact the patient, make an appointment if needed with the doctor, or send to ER	Medical assistant or nurse
Potentially urgent	Presence of suicidal thoughts	Contact the patient	Medical assistant or nurse
Needs action	Glucose at 16 mmol/l for 3 days	Make an appointment with doctor	Medical assistant or nurse
Highlight for next appointment	BP of 170/100 on a single occasion	To discuss at next planned visit	Nurse or doctor at next visit
Useful to know	Patient submitted a response or uploaded PGHD	Review when possible	Medical assistant or nurse
Absence of response	No response after a week	Resend an invitation for a questionnaire or to upload data	Medical assistant

These are examples of alert designs for PGHD, which can vary widely depending on the existing process of patient care in a division: some clinicians may want to be alerted directly for values that are less urgent. The heterogeneity in physicians' expectations and processes emphasize the need for customized design to ensure efficiency and patient safety.

4.2. Analysis of PGHD and alerts

Although PGHD can be represented as numerical data, thresholds for concern may vary from a person to another, even for a given parameter. Absolute values may suffice to trigger an alert for some parameters, while others need to assess the trend over time.

Table 2. Type of PGHD parameter and their modality

Parameter	Modality	Examples
PHQ-2* (clinical score to screen for depression)	1 threshold	If score ≥ 3, depression is likely
Pain scale (Visual analog scale)	Multiple thresholds	<4: mild pain 4<pain<6: moderate pain >6: severe pain
Specific question (e.g., suicidal thoughts)	Positive answer	Alert if "Yes"
Blood pressure (systolic)Δ	Multiple threshols	> 180 mmHg if single value >160 if sustained <90
Weight ¥	Upper threshold	Weight gain of ≥2 kg over 3 days
Glucoseα	Upper and lower threshold	Sustained values or recurrent pattern over ≥3 consecutive days (both > 15 mmol/l or < 3.5 mmol/l)

*Patient-health questionnaire

Δ Blood pressure values could be combined with symptoms to improve detection or pattern recognition (e.g., with chest pain)

¥ Example for individuals with heart failure Thresholds vary depending on target population

α Thresholds need to be adapted for older patients, or if on insulin or certain anti-diabetic medication

When the PGHD parameter has a binary interpretation (Y/N or one threshold), and does not require additional considerations, the alert depends on the degree of urgency. When the PGHD parameter has 2 thresholds or more, the urgency of the alert may vary depending on the threshold. Severe pain will be an urgent alert, while mild pain may be useful to consider at the next visit, for example. Yet other variables need to be assessed for patterns of concern, such as recurrent hypoglycemia over several days. Other considerations for these parameters are the target population, or concurrent elements (other symptoms, medications, or comorbidities, for example). PGHD of data that are interpreted as variations or trends can also use differential analyses or regression models, especially if the threshold is a constant (e.g., weight gain of 2 kg in 3 days). Examples of variables, modalities for analysis are presented in Table 2.

For PGHD parameters like blood glucose measurements, which typically result in many data points, or even continuous data from continuous glucose measurements are interpreted as patterns. In these cases, more advanced analyses using machine learning such as temporal analysis can be considered, and even adjusted by users who can confirm or infirm the results of the analyses.

4.3. Potential pitfalls

PGHD can result in vast amounts of data. Advanced models, such as machine learning systems can help to monitor and adapt alert thresholds, but should be use with caution given the possible consequences. While analyses are invaluable to avoid missing key results, overuse of analysis methods can lead to over diagnosis and alert fatigue. These would then lead to decreased efficiency for patient safety, with a possible increase in healthcare use (visits, tests and anxiety for patients).

5. Conclusion

Our exploration of approaches to integrate PGHD into the EMR emphasized the need to consider the type of data, which will guide the design of alert systems. These alerts must be implemented within local workflow processes, which will define the alert targets. Visual design needs to be adapted to each context, degree of urgency and distinction for process and content alerts. Finally, designing multi-patient views, and a simplified process to allow clinicians to specify which PGHD is needed for a given patient are also needed for all the participants interviewed.

References

[1] Shapiro M, Johnston D, Wald J, et al. Patient-Generated Health Data - White Paper. *Prep Off Policy Planning, Off Natl Coord Heal Inf Technol Res Triangle Park NC RTI Int* 2012; 35.

[2] Blondon K, Ehrler F. Integrating patient-generated health data in an electronic medical record: Stakeholders' perspectives. *Stud Health Technol Inform* 2020; 275: 12–16.

[3] Pearce A, Haas M, Viney R, et al. Incidence and severity of self-reported chemotherapy side effects in routine care: A prospective cohort study. *PLoS One* 2017; 12: e0184360.

Challenges of Trustable AI and Added-Value on Health
B. Séroussi et al. (Eds.)

doi:10.3233/SHTI220444

Impact of Electronic Medical Records Systems in Reporting HIV Health Data Indicators in Kenya

Philomena N. NGUGI[a,b,1], Milka B. GESICHO[a] and Ankica BABIC[a,c]
[a] *Department of Information Science and Media Studies, University of Bergen, Norway*
[b] *Institute of Biomedical informatics, Moi University, Kenya*
[c] *Department of Biomedical Engineering, Linköping University, Sweden*

Abstract. Electronic Medical Records Systems (EMRs) improve the quality of patient care and reduce medical errors. Nevertheless, their role in health data indicator reporting performance is unclear. We assessed reporting completeness and timeliness of HIV indicator data to the national aggregate reporting system, District Health Information Software 2 (DHIS2) in Kenya. We compared the reporting performance of facilities with and without EMRs implementation for the year 2013 as EMRs uptake was in progress. The comparative analysis involved 104 facilities implemented with and 152 without KenyaEMR system on three HIV programmatic areas. There were no statistically significant differences in performance regarding reporting completeness and timeliness by facilities with or without EMRs (p-values > 0.05 on all the three areas). The KenyaEMR system assessed in this study, therefore, cannot be associated with the transformed performance in reporting health indicators. This was probably due to the fact that the EMRs do not report electronically to DHIS2. Additional analysis can be conducted to compare reporting performance once data exchange functionality is fully established between KenyaEMR and DHIS2 systems.

Keywords. EMRs, reporting performance, completeness, timeliness

1. Introduction

Reporting health care data to support decision making is key to realization of global health goals especially in resource-limited settings towards achieving UNAIDS 95-95-95 global HIV epidemic control goals: 95 percent of people living with HIV know their HIV status, 95 percent of people who know their HIV status are accessing treatment, and 95 percent of people on treatment have suppressed viral loads [1,2]. The capability to efficiently report health indicators requires timely, reliable, high-quality and accessible health service data [3]. Introduction of Electronic Medical records systems (EMRs) in health care delivery has shown improvement in time dependent events such as patient waiting time, time to processing specimen in the laboratory from test request to results reporting among others benefits [4]–[6].
HIV related data reporting to the national reporting system, DHIS2, is a mandatory requirement in Kenya by the Ministry of Health in several programmatic areas. Health

[1] Corresponding Author, Philomena N. NGUGI, Department of Information Science and Media studies, University of Bergen, Norway; Email: waruharip@gmail.com.

indicator reporting forms in paper format were introduced in the year 2011 in all healthcare facilities in the 47 counties countrywide while DHIS2 system was rolled out in year 2012. The first Electronic Medical Record systems were introduced to support patient data management in the year 2012 but to a wider scale in 2013. The uptake of varied types of EMRs supported by different partners has been growing rapidly [7]–[9]. The purpose of this study was to assess immediate impact of EMRs in reporting health data indicators with respect to reporting completeness and timeliness by the various facilities across the counties. Completeness and timeliness components are key data quality attributes leading to quality of care.

2. Methods

HIV facility reporting data were extracted from the DHIS2 system for HIV Counselling and Testing (HTS), Prevention of Mother to Child Transmission (PMTCT) and Care and Treatment (C&T) programmatic areas. Systematic procedures were used in cleaning the data. The facility reporting data for these programmatic areas was merged using Master Facility List codes (unique identifier for facilities), with an updated list of facilities containing KenyaEMR system implementation dates. Only facilities extracted from DHIS2 that matched with those updated in the KenyaEMR implementation list were retained. The EMRs implementation year for facilities in the resulting database was used to segment facilities based on those with EMRs implementation and those without EMRs implementation in a particular year. For instance, a facility reporting to DHIS2 in the year 2012 may have had EMRs implemented in 2013, hence were categorized as having no EMRs implementation for the year 2012.

KenyaEMR system implementation in facilities commenced in the year 2012. Based on the KenyaEMR system implementation date list, there were only seven implementations in the year 2012, hence, analysis was limited to only year 2013 as there were more implementations. In addition, all facilities in our data set had EMRs implementations by 2014. Performance assessment was based on facility reporting completeness (percentage of expected reports submitted) and timeliness (percentage of expected reports submitted on time).

Prior to data analysis, normality tests we conducted using Shapiro-Wilk tests and test of Homogeneity of Variance. Consequently, we conducted Mann-Whitney U tests to compare the relationship between the facilities implemented with KenyaEMR and those without implementation to performance on HIV indicator reporting, given that the data was not normally distributed. All analysis was conducted using SPSS.

3. Results

On average, a total of 256 facilities qualified for the comparative analysis across the three programmatic areas. The comparative analysis resulted in p-values > 0.05 on all the three areas regardless of the state of implementation (Table 1). This reveals no statistically significant differences in performance in reporting completeness and timeliness for facilities with EMRs implementations and those without EMRs implementation in the various programmatic areas.

Table 1. Performance for completeness and timeliness in facilities with EMRs implementations and facilities without EMRs implementation for the year 2013

Quality attribute	Implementation Status	n	P-value	Mean Rank
	HTS			
Completeness	Without EMRs	152	0.236	124,47
	With EMRs	104		134,39
Timeliness	Without EMRS	152	0.296	132,44
	With EMRs	104		122,75
	PMTCT			
Completeness	Without EMRs	153	0.097	123,91
	With EMRs	105		137,65
Timeliness	Without EMRS	152	0.546	131,80
	With EMRs	104		126,15
	C&T			
Completeness	Without EMRs	152	0.186	123,37
	With EMRs	103		134,83
Timeliness	Without EMRS	152	0.549	130,26
	With EMRs	103		124,67

In Figure 1 and Figure 2, we illustrate the distribution of facilities with and without KenyaEMR implementation in the various counties. The aim was to have a visual representation of the reporting performances by facilities in the various counties. Nonetheless, the number of facilities is varying and limited to enable any analyses between the counties.

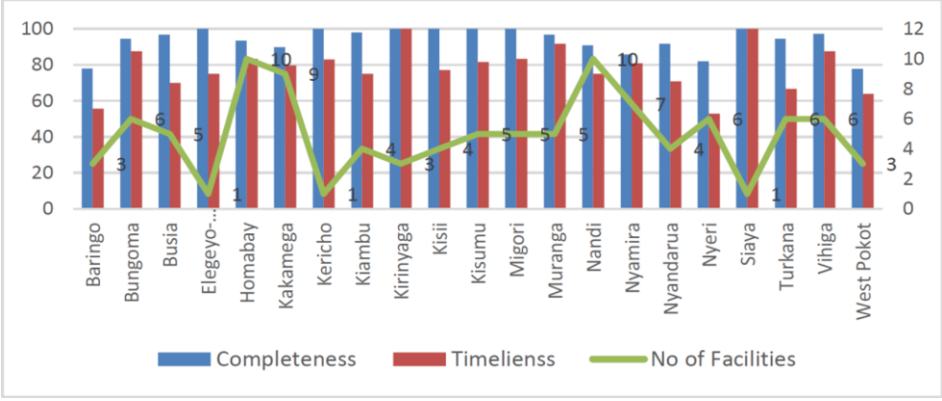


Figure 1. Percentage average reporting of HTS for facilities with EMRs implementation and their distribution in the various counties

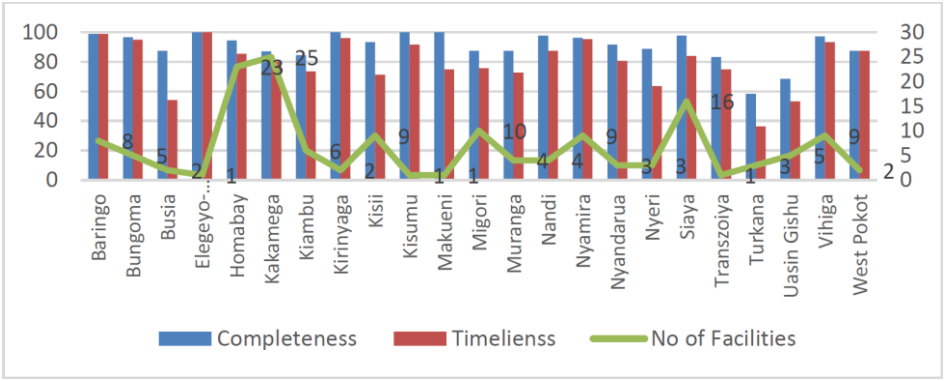


Figure 2. Percentage average reporting of HTS for facilities without EMRs and their distribution in the various counties

4. Discussion

As seen in Figures 1 and 2, performance on reporting completeness and timeliness varied within the counties and within facilities with and without EMRs implementation in the year under study. Nonetheless, our analysis was based on cumulatively all facilities with and without implementation, and not disaggregated by counties (Table 1). This enabled comparison of performance in facility reporting completeness and timeliness within the initial years of EMRs implementation. Ordinarily, it can be assumed that differences in reporting performance are expected in facilities with EMRs implementations verses those without. However, this study reveals that there were no differences in performance among facilities with EMRs and those without despite their contribution to improved internal health care services as reported by some studies [4]–[6]. This can be explained due to the fact that EMRs were not involved in direct reporting of the indicators to DHIS2 system.

While there are other types of EMRs implementations in Kenyan health facilities [10], the study was limited to only those implemented and projected to implement KenyaEMR system as it was the only one where implementation dates were provided. Additionally, the Kenyan MoH has adopted KenyaEMR as the national EMRs.

This study suggests that there was no direct effect on the reporting performance on introducing the KenyaEMR system which seemed to be contrary to the expectation. That could be attributed to the reporting routines, training and availability of the staff to transform and migrate data between the EMRs and DHIS2 systems. Nevertheless, this study can be used as a baseline for future comparisons in evaluating EMRs implementations in relation to health indicator reporting performance.

5. Conclusion

We found that the implementation of KenyaEMR system considered in this study cannot be associated with transformed performance in reporting HIV health data. Our next step is to conduct a follow-up study to investigate reporting performance as implementations have progressed including other types of EMRs as well as. Comparison of performance

after establishment of electronic health data exchange between the national reporting system, DHIS2 and EMRs will also be of interest in our future study.

Ethical Approval and Acknowledgements

Ethical approval was obtained from the Institutional Review and Ethics Committee in Moi University (MU/MTRH-IREC FAN:0003348). This work was supported in part by the NORHED program (Norad: Project QZA-0484). The content is solely the responsibility of the authors.

References

[1] PEPFAR, Monitoring , Evaluation , and Reporting Indicator Reference Guide. 2018 November; p 23–26.
[2] Castelnuovo B et al, Implementation of Provider-Based Electronic Medical Records and Improvement of the Quality of Data in a Large HIV Program in Sub-Saharan Africa. PLoS One. 2012 Dec;7(12):e51631.
[3] Kariuki JM et al. Automating indicator data reporting from health facility EMR to a national aggregate data system in Kenya: An Interoperability field-test using OpenMRS and DHIS2. Online J. Public Health Inform. 2018;8(2):e188.
[4] Westbrook JI, Georgiou A, Dimos A, Germanos T. Computerised pathology test order entry reduces laboratory turnaround times and influences tests ordered by hospital clinicians: A controlled before and after study. J. Clin. Pathol. 2006,59(5):533–536.
[5] Were MC et al. Evaluation of computer-generated reminders to improve CD4 laboratory monitoring in sub-Saharan Africa: A prospective comparative study. J. Am. Med. Informatics Assoc. 2011;18(2):150–155.
[6] Alamo ST et al. Electronic medical records and same day patient tracing improves clinic efficiency and adherence to appointments in a community based HIV/AIDS care program, in Uganda. AIDS Behav.2012;16(2):368–374.
[7] "KenyaEMR Implemented at More Than 340 Sites in Under Two Years – I-TECH." [Online]. Available: http://www.go2itech.org/2014/11/kenyaemr-implemented-at-more-than-340-sites-in-under-two-years/. [Accessed: 22-Nov-2017].
[8] "Palladium Group International," 2018. [Online]. Available: https://en.m.wikipedia.org/wiki/Palladium_International. [Accessed: 20-Oct-2018].
[9] Kang'a S et al. A national standards-based assessment on functionality of electronic medical records systems used in Kenyan public-Sector health facilities. Int. J. Med. Inform.2017;97:68–75.
[10] Powell J, Paton C. Survey of Electronic Health Record (EHR) Systems in Kenyan Public Hospitals : A mixed-methods survey. JMIR Med. informatics. 2019 December.

Challenges of Trustable AI and Added-Value on Health
B. Séroussi et al. (Eds.)

doi:10.3233/SHTI220445

Collecting Data from a Mobile App and a Smartwatch Supports Treatment of Schizophrenia and Bipolar Disorder

Steinunn Gróa SIGURÐARDÓTTIR [a,1], Anna Sigridur ISLIND [a] and María ÓSKARSDÓTTIR [a]

[a] *Reykjavik University, Department of Computer Science, Reykjavik, Iceland*

Abstract. Mental disorders affect individuals and societies around the world negatively, with the health-related burden of 32,4% out of the overall disease burden. This large part of the overall burden underlines a growing need for innovation to support the treatment of mental disorders like schizophrenia and bipolar disorder. This empirical study features two groups of patients; a group of nine patients diagnosed with bipolar disorder and a group of twelve patients diagnosed with schizophrenia. The patients in the study carry a smartwatch for six weeks, continuously collecting data into a digital health platform. Additionally, they answer five daily wellbeing questions in a mobile app. To supplement that data, they also answer a questionnaire three times over the interval and at the end of the period they attend a semi-structured interview. We offer four main aspects to consider for PGHD in mental health: i) sharing data easily with healthcare professionals, ii) being able to engage with your own PGHD, iii) the watch use can help the patients regulate routine in their daily life, iv) tonality and phrasing.

Keywords. Schizophrenia, Bipolar Disorder, Digital Platform, mHealth, eHealth, Smartwatch, Mobile Application, Mental Health

1. Introduction

Mental disorders cause individuals and societies around the world effects that cannot be measured easily. The estimated mental health-related burden accounts for 32.4% of years lived with disability (YLD) of all YLDs' in the world [1], not to mention the often overlooked effects mental illness can have on close family and friends. One of the more severe, chronic mental disorders is bipolar disorder (BD), which has a reported prevalence of 0.6% [2]. Patients diagnosed with BD experience extreme mood swings and activity fluctuations from being hyperactive to total inactivity. Patients with BD generally suffer from sleeping difficulties and may struggle with day-to-day tasks [3]. Schizophrenia is another serious chronic mental disorder with a reported prevalence of 0.3% [2]. The disorder causes extensive paranoia and delusions that affect the quality of life negatively. Schizophrenia patients also face various life challenges, such as a low employment rate (below 20%) and high homelessness (up to 20%) [4]. Research suggests that increased

[1] Corresponding author: Steinunn Gróa Sigurðardóttir; e-mail: steinunngroa@ru.is.

physical activity can improve wellbeing for patients with schizophrenia as well as BD [5]. In fact, these two disorders are often researched together due to similarities in symptoms [6].

Digital platform (hereafter "platform") is essentially software that facilitates a connection between user's needs and what a service provider has to offer. The supply and use of platforms is becoming increasingly established, for example in healthcare [7]. In a healthcare setting, a platform can potentially create a bridge between a patient and healthcare professionals through monitoring of patient-generated health data (PGHD). PGHD in a platform setting encompasses data produced and collected by patients, which is brought into a healthcare platform to enhance the quality of care. PGHD is any clinically relevant data collected and used by patients and either shared or not shared with healthcare professionals, and PGHD can in some sense contribute to increased wellbeing or health outcome. A study on PGHD from wearables for self-monitoring reports that patients believed that using wearables to regulate their symptoms has the potential of improving the overview of their condition [8]. In this project we collect PGHD with smart devices; both automatically (with a smartwatch) and manually via a mobile app. All data is collected and visualized in a platform.

This paper is based on an empirical study that features two groups of care recipients (hereinafter patients) who are under medical care; one group diagnosed with BD and the other group with schizophrenia. The interventions with the patients include increased physical activity, which has proven effective for treatment of the two disorders. Patients also attend structured consultations with a healthcare professional where they are asked, among other things, to assess their quality of sleep. Our project entails the design of a platform that enables health data to be visualized and accessed, to involve the patients to a larger extent in their own care. Through the empirical, multidisciplinary research project reported on herein, our goal is to provide insights into both the clinical value as well as the increasingly important value of PGHD in mental healthcare. From that, we seek to answer the research questions: i) How should a smartwatch and a mobile app be used to support treatment of schizophrenia and BD? and, ii) how can a digital a platform function as a bridge between patients and healthcare professionals?

2. Methods

We have designed and developed a digital health platform to be used in psychiatric care, called DataWell; which reflects the use of data for wellbeing. The platform is under further development through a co-design approach in this project. Our platform combines the following PGHD types. From the a Withings Steel HR smartwatch we gather heart rate monitoring (beats per minute), sleep monitoring (sleep duration, sleep depth, interruptions), activity monitoring (estimation of ten activity types), and step count (based on distance). The sensors in question are: heart rate infrared sensor, day and night motion sensor and high precision MEMS 3-axis accelerometer [2]. A reason for choosing this particular smartwatch is their unusually low power consumption. The battery charge lasts up to 25 days, which we consider an important factor when it comes to patients with severe mental illnesses. Additionally, Withings provide accessible, structured data through an API, which is convenient when building a new platform. Although outside the scope

[2] https://www.withings.com/eu/en/steel-hr

of this paper, it is important to mention that we have data from the electronic patients records (EPR/EHR) for two years.

The process of the data gathering is as follows. The patients who accept to partake in this six weeks study answer a questionnaire three times (week one, three and six). The questionnaire includes a collection of validated constructs to measure level of anxiety, depression severity, view towards technology, self-efficacy, empowerment, and impact of self-monitoring. The idea behind that data collection is to create a link between the outcome of the questionnaire and the way patients choose to use the platform, watch and mobile app. They carry a smartwatch for these six weeks, collecting data into our platform. They also answer five daily questions on general wellbeing in our mobile app, which we designed as a part of this project and the mobile app feeds data into the platform. During the research period, the healthcare professionals are able to monitor their patients, through the platform. Both patients and healthcare professionals attend semi-structured interviews separately at the end of the six-week period, where the focus is on user experience and usability of the platform and their view on data-driven mental health and on continuously collecting PGHD. The analytical approach for the interviews is content analysis and the results will be used to further guide the design and development of the platform. In this paper, we focus on findings from the patients. We have included 21 patients in total. Three of those dropped out during data collection due to sickness or issues related to comfort. From that, nine patients are diagnosed and treated for BD (two dropped out) and twelve patients are diagnosed and treated for schizophrenia (one dropped out). Thus, seven patients with BD and eleven with schizophrenia have completed all data collections steps.

3. Results

The majority of patients felt encouraged by monitoring their data continuously and expressed interest in continuing the data collection. Meeting the patients three times over the data gathering interval, and especially the in-depth interview at the end has enabled a unique opportunity to understand the patients and how this approach affects them. One patient summarized her view with the words: “This really matters.” which is encouraging for the continuation of the study.

Overall, we identify four main findings. *Firstly,* sharing data with healthcare professionals seems to come quite easily to them. Most of the patients are used to sharing all sorts of things with healthcare professionals and feel that the data sharing in this experiment is no invasion, but rather supportive. On that note, another patient shared: “It’s good that healthcare professionals have access to the data, because they’re just trying to help you.”

Secondly, an important finding has to do with the empowerment of the patient, a feeling that many of them describe. One patient said: “Seeing your own health data is a good feeling.”, while another shared: “It encourages you to see an overview of the steps.” What can be learned from that is that observing and monitoring one’s own data helps in the empowerment process.

Thirdly, is that the usage of the watches is different between individuals. One patient described his opinion on how the experience with the watch has been like: “I’m not quite sure, but it was really new to me because I have not worn anything like this before that

measures heart rate, steps and sleep." Continuing, now on his view on if there have been some changes since the start of data gathering: "Yeah, a little bit at least, I'm going to bed at the right time now." In general, some patients like to monitor their data closely, see the steps "coming in" over the course of the day and reflect on them in terms of how they feel and how they're doing that day. Others want to let it rest for a few days and then see a chunk of data in the platform. This gives them context and understanding on they're feelings and the development of the disease for the days in question.

Lastly, regarding the mobile app and the way that the questions in there are formulated, we found that inclusiveness in all communication with patients and putting ourselves in the patients' shoes is truly important. One patient said: "Maybe what I got out of this was that you just reflect a little bit about how, you know you get these questions, then you wonder how was the stress today. You tune a little bit into yourself with this too." Another patient expressed gratitude and satisfaction in the phrasing of the questions in the mobile app stating that the questions were inclusive and that they addressed them as a "thinking human being" and not a "7 year old" or "someone stupid."

4. Discussion

When discussing health platforms and PGHD, it is vital to reflect on the data collected, the way the data is presented and used. To begin with, data can consist of either a few values of data points or multiple ones and the granularity of data is thereby important to consider [9]. Also, when showing trends, they can be relational, hierarchical, or a combination of the two, which is an important aspect to consider when choosing types of graphs that can help the patient understand their data. Each data entry is not an autonomous entity; rather, it is a part of an array in which data over time builds up a larger data set that can then be visualized, for the dual purpose of i) functioning as decision-support for the healthcare professionals and ii) triggering self-care and self-monitoring for patients [10]. We consider self-monitoring to be the tracking of health-related information for patients and using that information to monitor wellbeing or health systematically. Prior research has shown that using wearables for self-monitoring can introduce an increase in the activity of the consumer [11], and in this paper, we corroborate those findings by showing that wearables are by most in our patient group, considered empowering. With that in mind, when designing a health platform that takes in extensive amounts of data, we recommend that it is done in a participatory work with all stakeholders involved and especially highlight the importance of giving the patients a voice in that process. Cognitive impairment is a common observation for individuals with schizophrenia [6] and BD [12], it can result in difficulties with written information. Because of that, the way PGHD enters the conversation, and the way the data is collected, discussed and used are key aspects of this research project. Consequently, we focused on helping the patients collect the PGHD, to understand their data and to reflect on their data. From our findings we conclude that in mental healthcare, PGHD can improve the wellbeing of patients. PGHD points towards a patient perspective through patient centricity and PGHD can improve partnership with the healthcare professionals, in line with [13].

In conclusion we have derived knowledge about the kind of data that is useful in mental health services and when it proves to be an effective addition to existing clinical treatment. Our findings can be summarized and forwarded through four major aspects: i)

sharing data easily with healthcare professionals is an important factor for a platform to be implemented in clinical practice, ii) being able to engage with own PGHD, empowers patients that are diagnosed with BD or schizophrenia, iii) the watch use can help the patients regulate routine in their daily life, iv) tonality and the way patients that are diagnosed with BD or schizophrenia are addressed in conversations, is vital. From our findings, we conclude that collecting PGHD into a platform outlines a key feature of future support in treatment of schizophrenia and BD.

4.1. Limitations

In this project the patients observe their own physiological health data, combined with their daily mental status, to observe trends in how their psychological health fluctuates in context with the physiological data signs. However, some patients forgot to fill out the daily survey in the mobile app, which might have caused them overlook those trends. Future work will include notifications in the mobile app, to compensate for that. Another direction in future work is to analyze if some patients show signs of dis-empowerment through an approach like this, possibly with a higher level of anxiety, stress, dropout rate, etc. Another aspect worth mentioning, is the inclusion of patients which was a considerable challenge due to their preexisting diagnosis, which renders them unable to predict what kind of day it will be; if they will be able to show up or not. There was about a 50% dropout rate during the appointments, but with perseverance of the first author when re-booking sessions multiple times, we managed to get all patients to meet with us.

References

[1] Vigo D, Thornicroft G, Atun R. Estimating the true global burden of mental illness. The Lancet Psychiatry. 2016;3(2):171-8.
[2] Ritchie H, Roser M. Mental health. Our world in data. Retrieved May. 2018;19:2020.
[3] Mansell W, Morrison AP, Reid G, Lowens I, Tai S. The interpretation of, and responses to, changes in internal states: an integrative cognitive model of mood swings and bipolar disorders. Behavioural and Cognitive psychotherapy. 2007;35(5):515-39.
[4] McNiel DE, Binder RL, Robinson JC. Incarceration associated with homelessness, mental disorder, and co-occurring substance abuse. Psychiatric Services. 2005;56(7):840-6.
[5] Callaghan P. Exercise: a neglected intervention in mental health care? Journal of psychiatric and mental health nursing. 2004;11(4):476-83.
[6] Bowie CR, Harvey PD. Cognitive deficits and functional outcome in schizophrenia. Neuropsychiatric disease and treatment. 2006;2(4):531.
[7] Islind AS. Platformization: co-designing digital platforms in practice. University West; 2018.
[8] Wu RC, Ginsburg S, Son T, Gershon AS. Using wearables and self-management apps in patients with COPD: a qualitative study. ERJ open research. 2019;5(3).
[9] Perin C, Vuillemot R, Stolper CD, Stasko JT, Wood J, Carpendale S. State of the art of sports data visualization. In: Computer Graphics Forum. vol. 37. Wiley Online Library; 2018. p. 663-86.
[10] Islind AS, Lindroth T, Lundin J, Steineck G. Shift in translations: Data work with patient-generated health data in clinical practice. Health informatics journal. 2019;25(3):577-86.
[11] Brickwood KJ, Watson G, O'Brien J, Williams AD. Consumer-Based Wearable Activity Trackers Increase Physical Activity Participation: Systematic Review and Meta-Analysis. JMIR mHealth and uHealth. 2019;7(4):e11819.
[12] Lima IM, Peckham AD, Johnson SL. Cognitive deficits in bipolar disorders: Implications for emotion. Clinical psychology review. 2018;59:126-36.
[13] Burns K, McBride CA, Patel B, FitzGerald G, Mathews S, Drennan J. Creating Patient Generated Health Data: Interviews and a Pilot Trial Exploring How and Why Patients Engage.

Challenges of Trustable AI and Added-Value on Health
B. Séroussi et al. (Eds.)

doi:10.3233/SHTI220446

Data Ingestion for AI in Prostate Cancer

Haridimos KONDYLAKIS[a], Stelios SFAKIANAKIS[a], Varvara KALOKYRI[a], Nikolaos TACHOS[a], Dimitrios FOTIADIS[a], Kostas MARIAS[a] and Manolis TSIKNAKIS[a]
[a] *FORTH-ICS, N Plastira 100, Heraklion, Crete, Greece*

Abstract. Prostate cancer (PCa) is one of the most prevalent cancers in the male population. Current clinical practices lead to overdiagnosis and overtreatment necessitating more effective tools for improving diagnosis, thus the quality of life of patients. Recent advances in infrastructure, computing power and artificial intelligence enable the collection of tremendous amounts of clinical and imaging data that could assist towards this end. ProCAncer-I project aims to develop an AI platform integrating imaging data and models and hosting the largest collection of PCa (mp)MRI, anonymized image data worldwide. In this paper, we present an overview of the overall architecture focusing on the data ingestion part of the platform. We describe the workflow followed for uploading the data and the main repositories for storing imaging data, clinical data and their corresponding metadata.

Keywords. Data Ingestion, Prostate Cancer

1. Introduction

About 1,300,000 citizens of the European Union are estimated to have had a prostate cancer diagnosis in the last five years [1]. The severe socioeconomic burden for health services and the negative effects on the quality of life of patients call for immediate actions [2][3]. Artificial Intelligence on the other hand has the potential to bring medicine from the era of 'sick care' to the era of healthcare and prevention, fueled by the availability of large datasets ("big data"), substantial advances in computing power, and new deep-learning algorithms [4]. However, the availability of large, quality-controlled datasets for building those AI models, currently remains a major challenge. To this purpose, several health imaging repositories have been created [5][6], such as the Cancer Imaging Archive (TCIA) [7]. However, the vast majority of these repositories have been created as stand-alone entities, being currently not in a position to become interoperable with similar existing initiatives. The need for the creation of a fully FAIR (Findable, Accessible, Interoperable, Reusable), GDPR compliant, European imaging repository still stands [8] and has been recognized by other EU research projects like PRIMAGE [9] and CHAIMELEON [10] however, still there are not tangible results from these projects. ProCAncer-I's vision is to become a catalyst in this process by creating the first European, ethical- and GDPR (General Data Protection Regulation) compliant, quality-controlled, prostate cancer (PCa) related, medical imaging platform, in which both large-scale data and AI algorithms will co-exist. To this end, the ProstateNet dataset featuring an unprecedented 1.5 million image representations of prostate cancer will be created within a sustainable AI cloud-based platform for the development, implementation, verification and validation of trustworthy, usable and reliable AI models. In this paper

we provide a glimpse of the overall architecture focusing on the data ingestion part of the platform. model used.

2. Architecture

ProCAncer-I aims to deliver an infrastructure that follows the principles of open source, FAIR data access, common look-n-feel, common authentication and authorization, layered developing of modelling service, modelling service certification and cloud infrastructure independence. The logical view of the ProCAncer-I platform with the main domain specific areas of functionality of the system is shown in Figure 1. The following subsystems are identified:

Data ingestion and upload. This includes all the infrastructure (tools, services, cloud resources) that allows a data provider to upload their data sets according to the project's guidelines and best practices (e.g. anonymization) so that they become integrated to the curated cancer-related data managed by the system.

Data Management, which supports the "data at rest" scenarios, is the core of the platform supporting all the other subsystems for the storage, efficient indexing, curation, and retrieval of the data.

Domain specific tools, for example image and data annotation and data *tools*, which support domain experts to annotate and curate the imaging data.

Model management. This is the part of the platform supporting the management of computational and AI tools and models. It allows searching for available models, the development of new ones, model execution and monitoring, etc.

Data and Service "Peering" tools, that support the exchange of data and interoperability of services with other research infrastructures using well defined FAIR-enabled APIs and applications like the "Honest Broker".

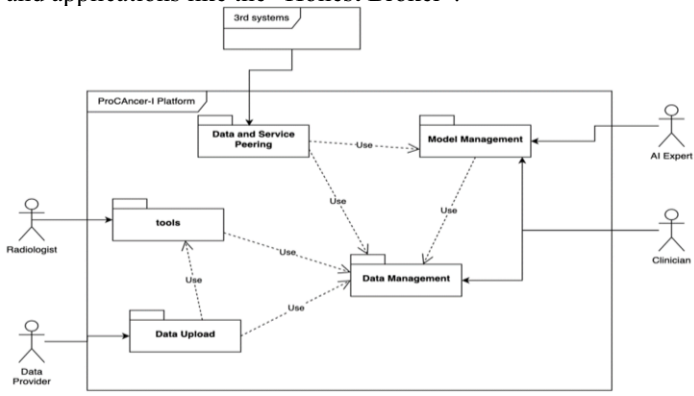


Figure 1. The main subsystems of the ProCAncer-I platform

Hence, during the initial implementation period the main focus has been on the design, development and delivery of the infrastructure and tools to enable data collection and its preparation, including de-identification, for upload into the platform. In the sequel we present data ingestion and upload, data management and domain specific tools that have been developed and integrated to allow data providers to make their data sets available to the ProCAncer-I community securely and fully annotated.

Data ingestion and upload. The ProCAncer-I platform will collect and manage large amounts of multimodal data (mpMRI imaging data and related clinical data) and

metadata to be used for the training of advanced AI models. The ProCAncer-I cloud platform storage, ProstateNet, is comprised of 3 components: a) the DICOM Object Store which stores medical imaging data; b) the clinical data document store which stores the clinical data; and c) the meta-data catalogue which stores metadata and semantic annotations to enable rich search and discovery of data and its exploitation.

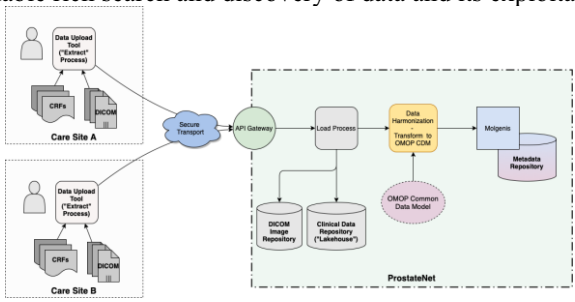


Figure 2. Main components and processes of the data ingestion pipeline

The whole data ingestion pipeline with its main components and subprocesses is shown in Figure 2. The clinical partners use a local, integrated eCRF and data upload tool to organise the DICOM studies and complete the clinical information, validate the use case, anonymise data and upload data to the cloud staging area. Each Clinical Partner is able to run the data curation tools (if needed), verify the anonymisation and completeness of data, and submit validated cases to the ProstateNet repository (so called "staging area").

DICOM Image Repository. The ProCAncer-I project DICOM Image Repository provides the necessary services for saving, updating, and retrieving DICOM studies. The implementation of the repository is compliant with both the DICOM and HL7 standards, thus allowing seamless interoperability with existing PACS systems and scanners. To support the several steps of data curation, annotation and AI research and development, the repository provides mechanisms for querying and retrieving data through an API gateway using the standard DICOMWeb [11] suite of programming interfaces.

Clinical Data Repository. The Clinical Data Repository is the central data warehouse of the ProCAncer-I platform, accessible through a web-based ("RESTful") API. In the data upload process this repository stores the submitted clinical and image related information before the metadata extraction and their persistence in the Metadata Catalogue. In this way, the data ingestion follows an Extract-Load-Transform (ELT) [12] design where this repository is responsible for maintaining the data in their original submitted format. The Clinical Repository is similar to a "data lake" that contains all of the uploaded information (except the imaging data) in the format that was uploaded. Being similar to a traditional data warehouse it offers transactional interactions and both structured and unstructured data storage. In addition, it contains imaging related metadata, for example selected DICOM tags extracted from the uploaded data, in a quasi-relational schema so that complex queries are possible and efficient. In contrast to the DICOM Imaging Repository that offers a mostly key-based "blob" storage (e.g. retrieving a DICOM series by its series UID), the Clinical Repository is able to cope with lots of different search criteria and access patterns. It also stores information about the results of the image processing tools (e.g. segmentations) and their parameters allowing the linking between imaging data and provenance related metadata extraction. Finally, it maintains an upload log so to enable traceability, by gathering of statistics and monitoring of the use of the data platform are possible.

Meta-Data Catalogue. In ProCAncer-I, the MOLGENIS (https://www.molgenis.org/) platform has been adopted to serve as the main metadata catalogue of the project, whereas OMOP-CDM and its extensions are used as the common data model to store the available metadata. MOLGENIS has a completely customized data system allowing modeling of the data using external data models. In addition, it is modular, having several modules to store and interact with the stored data, and provides interfaces to create R and Python scripts that interact with the data. MOLGENIS takes away the hassle of storing data, and makes it highly accessible with filters and fast search capabilities.

Data Models. ProCAncer-I adopts the OMOP-CDM, which is one of the most widely used common data models for supporting analysis of observational health data. It supports the standardization and harmonization of health data as well as the generation of reliable scientific evidence about disease history, effects of medical interventions and health care interventions and outcomes. Besides the standard CDM, OMOP-CDM extensions are used, such as the Oncology CDM extension for representing cancer data at the levels of granularity and abstraction required to support cancer research. For radiology exams, although those can be currently registered using the OMOP-CDM, the model does not enable the storage of the subsequent curation process. As such, the ProCAncer-I aspires to introduce a radiology extension and is currently working on it in collaboration with the OHDSI Medical Imaging Working Group, focusing on including annotation, segmentation and curation data as radiomics.

OMOP-CDM ETL. To get from the native/raw data provided by the clinical sites to the OMOP Common Data Model (CDM), an extract, transform, and load (ETL) process was defined and implemented. This process transforms the data from its initial raw format to the CDM, and adds mappings to a set of Standardized Vocabularies. Terms found in the source data are mapped to concepts in the OMOP standard vocabularies to achieve semantic interoperability. In most cases a mapping to a standard concept with the same meaning as the source term can be made. If this is not possible, the source term is mapped to a non-standard concept. If a non-standard concept matching the source term does not exist either, then we create a custom 'ProCAncerI' concept. Concerning the radiology image metadata accompanying each one of the use cases, we have designed and implemented an initial CDM-extension (based on the current R-CDM extension proposed by the OHDSI community) for storing all image related metadata required by the project. In addition, we have designed an initial schema for storing the image curation information. During the whole pipeline the users are able to observe the various steps and specific tools for quality control are available. These tools include visualization interfaces for the imaging data and segmentation masks, as well as for the ETL output and the OMOP-CDM mapping and storage in MOLGENIS.

Deployment. Apart from the functional requirements and the data ingestion pipeline presented above, the envisaged platform needs to address certain non-functional requirements, such as scalability, security, availability, and overall performance. To address these needs, the platform is deployed on a commercial cloud using "cloud-native" technologies and tools. In particular, all the platform's components are deployed as containers using Kubernetes as the container orchestration platform for automated container deployment. The platform network is also "containerized" with both control and data plane composed of microservices with flexible deployment specifications to address fluctuations in workload. Finally, there's built-in observability and analytics functionality in order to enable continuous monitoring and automated troubleshooting

during the upload and data transformation phases, using existing monitoring tools like Prometheus (https://prometheus.io/) and Grafana (https://grafana.com/).

3. Conclusions

This paper provided an overview of the initial version of the ProCAncer-I Platform data ingestion modules. The various storage modules were described, including the DICOM Image Repository, which is compatible with the DICOM and DICOMweb standards, the Clinical Data Repository and the Meta-Data Catalogue, which is built on top of the highly customizable MOLGENIS application. Based on all of the above, the alpha version of the ProCAncer-I Platform is ready to securely accept the upload of retrospective anonymized data, pending the completion of integration of more functionalities. As next steps the data ingestion platform will be extensively evaluated by the users and also the ecosystem of services/tools comprising the AI framework is currently designed and under implementation. Those will be reported in a follow-up paper.

Acknowledgements

Research reported in this paper has been supported from the ProCAncer-I H2020 project (No 952159) and has received funding from the European Union's Horizon 2020 research and innovation programme.

References

[1] Epidemiology of prostate cancer in Europe: https://ec.europa.eu/jrc/en/publication/epidemiology-prostate-cancer-europe.
[2] Kondylakis H, et al. Patient empowerment for cancer patients through a novel ICT infrastructure. Journal of biomedical informatics 101 (2020): 103342.
[3] Kondylakis H, et al. Developing a data infrastructure for enabling breast cancer women to BOUNCE back. IEEE 32nd International Symposium on Computer-Based Medical Systems (CBMS). IEEE, 2019.
[4] Thrall JH, et al. Artificial Intelligence and Machine Learning in Radiology: Opportunities, Challenges, Pitfalls, and Criteria for Success, Journal of the American College of Radiology, 2018; 15(3):504-508.
[5] Sudlow C, et al. UK Biobank: An Open Access Resource for Identifying the Causes of a Wide Range of Complex Diseases of Middle and Old Age. PLoS Med (2015) 12:e1001779.
[6] Bamberg F, et al. Whole-Body MR Imaging in the German National Cohort: Rationale, Design, and Technical Background. Radiology (2015) 277:206–20.
[7] Clark K, et al. The Cancer Imaging Archive (TCIA): Maintaining and Operating a Public Information Repository. J Digit Imaging (2013) 26:1045–57. doi: 10.1007/s10278-013-9622-7
[8] Gabelloni M, et al. Bridging Gaps Between Images and Data: A Systematic Update on Imaging Biobanks. Eur Radiol (2021). doi: 10.1007/s00330-021-08431-6
[9] Martí-Bonmatí L, et al. PRIMAGE Project: Predictive In Silico Multiscale Analytics to Support Childhood Cancer Personalised Evaluation Empowered by Imaging Biomarkers. Eur Radiol Exp (2020) 4:22.
[10] Bonmatí LM, et al. CHAIMELEON Project: Creation of a Pan-European Repository of Health Imaging Data for the Development of AI-Powered Cancer Management Tools. Frontiers in Oncology. 2022:515.
[11] Genereaux BW, et al. DICOMweb™: Background and Application of the Web Standard for Medical Imaging. J Digit Imaging. 2018 Jun;31(3):321–326.
[12] Cohen J, Dolan B, Dunlap M, Hellerstein JM, Welton C. MAD skills: new analysis practices for big data. Proc VLDB Endow. 2009 Aug;2(2):1481–1492.

Challenges of Trustable AI and Added-Value on Health
B. Séroussi et al. (Eds.)

doi:10.3233/SHTI220447

Automated Coding in Case Mix Databases of Bacterial Infections Based on Antimicrobial Susceptibility Test Results

Radia SPIGA[a,b], François-Elie CALVIER[a], Anne CARRICAJO[c], Bruno POZZETTO[d], Béatrice TROMBERT-PAVIOT[a] and Cédric BOUSQUET[a,e, 1]
[a]*Public health and medical information unit, Saint Etienne University Hospital, France*
[b]*Mines Saint-Etienne, Univ Clermont Auvergne, CNRS, UMR 6158 LIMOS, Institut Henri Fayol, F – 42023*
[c]*Laboratoire des Agents Infectieux et d'Hygiène, Centre Hospitalo-Universitaire de Saint-Etienne, Saint-Etienne, France*
[d]*CIRI (Centre International de Recherche en Infectiologie), Université de Lyon, Université Claude Bernard Lyon 1, INSERM U1111, CNRS, UMR5308, ENS Lyon, Université Jean Monnet de Saint-Etienne, Lyon, France.*
[e]*Sorbonne Université, INSERM, Univ Paris 13, Laboratoire d'Informatique Médicale et d'Ingénierie des Connaissances pour la e-Santé, LIMICS, 75006, Paris, France*

Abstract. Our objective was to improve the accuracy of bacteria and resistance coding in a hospital case mix database. Data sources consisted of 50,074 files on bacteriological susceptibility tests transmitted with the HPRIM protocol from laboratory management system to electronic health record of the University hospital of Saint Etienne in July 2017. An algorithm was implemented to detect susceptibility tests containing information corresponding to codes whose addition in the case mix database was susceptible to increase the severity level of a diagnosis related group. Among 132 hospital stays fulfilling the conditions, 27 were lacking bacteria and/or resistance codes, and the tariff was increased for 9 stays, with earnings of €54,612. Analyzing Antimicrobial susceptibility tests helps to improve clinical coding and optimize the financial gain.

Keywords. Clinical Coding, Electronic Health Records, Bacteria, Health care payment system, France

1. Introduction

Funding of acute care in French hospitals is based on Prospective Payment System (PPS) that relies on the description of care activity. It is evaluated through a Diagnosis Related Group (DRG)-based information system called PMSI (*Programme de Médicalisation du Système d'Information* for "medicalized information system program"). This program uses the 10th revision of the International statistical Classification of Diseases and related health problems (ICD-10). A Patient Classification System is applied on the PMSI data to assign a hospital stay to a DRG. Each DRG has a tariff set by the Ministry of Health

[1] Corresponding Author, Dr Cedric Bousquet, SSPIM, Bâtiment CIM42, chemin de la Marandière, Hôpital Nord, 42055 Saint Etienne, France; E-mail: cedric.bousquet@chu-st-etienne.fr.

that varies according to the most responsible diagnosis and the severity scale from 1 to 4, corresponding to the severity of Comorbidities or Complications (CCs).

The PMSI database (French case-mix database) is commonly used for epidemiologic research and surveillance. However, several studies have pointed the low accuracy of associated diagnoses in the different hospital case-mix databases available worldwide, especially in case of bacterial infection [1–4]. In France, codes specifying the bacterium responsible for an infection or its resistance to certain drugs belong to the list of the CCs whereas codes related to highly-resistant bacteria induce an increased level of severity to 3 or 4.

We present here a pilot study using files sent from laboratory management system (LMS) to patients' electronic health record (HER). This file exchange follows the French HPRIM protocol (*Harmoniser et PRomouvoir l'Informatique* Médicale for "standardize and promote medical information technology") which allows a secure electronic transmission of biological results. We developed an algorithm for analyzing HPRIM files and targeting those responding to established criteria with the objective of demonstrating the feasibility to improve the coding of bacteria and resistance in a case mix database.

2. Methods

The results of susceptibility tests of bacteria from the Department of Infectious Agents and Hygiene of the University-Hospital of Saint-Etienne are stored in HPRIM files together with other data including the sampling site, the type of culture and the bacterium species. HPRIM files produced during the month of July 2017 by the LMS were extracted and stored safely on a server. At the time of extraction all the infectious agents were processed, including bacteria, viruses and parasites, without distinction.

```
Data: HPRIM document folder
Result: set of HPRIM documents containing epidemic alert
for all HPRIM documents in the folder do
    while not at end of this document do
        read file;
        if in OBX section then
            if results are complete and valid then
                bacteria name extraction from the OBX section;
                if bacteria in alert list then
                    if bacteria in presence alert list then
                        set alert flag on the document;
                    else
                        antibiogram extraction from the ATBL subsection;
                        if bacteria in multi-resistance alert list then
                            if bacteria resistance match multi-resistance
                             alert then
                                set alert flag on the document;
                            end
                        end
                    end
                end
            end
        end
    end
end
```

Algorithm 1: HPRIM extraction algorithm

Figure 1. Algorithm implemented to extract susceptibility tests from HPRIM files

We identified the parts of HPRIM format that describe antimicrobial susceptibility tests. We implemented a program using the Java language for reading the HPRIM files and selecting susceptibility tests related to bacteria and corresponding to a valid result, which excluded non-bacterial agents and invalid susceptibility tests figure 1). Bacteriological results corresponding to short hospital stays (less than 2 nights) were also excluded.

Then we targeted bacteria and resistance profiles corresponding to those whose presence leads to increase the severity level of the DRG (codes with severity levels of 3 and 4): the selected bacteria were *Staphylococcus aureus*, *Proteus mirabilis*, *Pseudomonas aeruginosa*, *Bacillus fragilis*, *Clostridium difficile* and *Clostridium perfringens*; the selected resistance profiles were methicillin-resistant *Staphylococcus aureus* (MRSA), extended-spectrum beta-lactamase-producing *Enterobacteriaceae* (ESBL-PE), vancomycin-resistant Enterococci (VRE) and carbapenemase-producing *Enterobacteriaceae*.

To code validly bacteria and antimicrobial resistance, the following PMSI regulatory conditions must be fulfilled: presence of a susceptibility test, mention of the infection on the EHR, and prescription of an antimicrobial treatment. For every hospital stay containing a susceptibility test according to the selection criteria, the accuracy of the ICD-10 coding in the hospital case-mix database was verified, with two possible scenarios: (i) in the absence of ICD-10 codes of bacteria or resistance, the EHR was checked to find information allowing the coding update; (ii) in its presence, the EHR was verified to find arguments justifying the coding. Then the sensitivity and specificity of initial bacterium coding were calculated. Finally, when missing ICD-10 codes of bacteria and/or resistance were added, the financial gain corresponding to the coding modification was calculated.

3. Results

From the 50,074 laboratory results performed by the laboratory on infectious agents during the month of July 2017, 159 susceptibility tests fulfilling the selection criteria were realized, corresponding to 132 hospital stays (Figure 2). After checking the case-mix database and the EHR, bacteria and antimicrobial resistance were appropriately coded for 20 hospital stays.

For the remaining stays without bacteria and/or resistance codes, conditions for adding such code were present for 27 stays (Table 1). Initial coding of bacteria and/or resistance in the PMSI database had a sensitivity of 42.5% (95% CI 28.26 - 57.82) and a specificity of 97.6% (95% CI 91.76 - 99.71). Among the 27 hospital stays that fulfilled an upgrading of coding conditions, the tariff of 9 stays was significantly increased, with overall earnings of €54,612. Among stays with unfulfilled coding conditions, two stays were wrongly coded; codes were removed without financial consequence.

Table 1. Comparative results of two PMSI coding strategies relative to 132 hospital stays including a significant bacterial infection with susceptibility testing, recorded in July 2017 at the University-Hospital of Saint-Etienne, France.

	Retrospective coding: Infection	Retrospective coding: No Infection	Total
Initial coding: Infection	20	2	22
Initial coding: No infection	27	83	110
Total	47	85	132

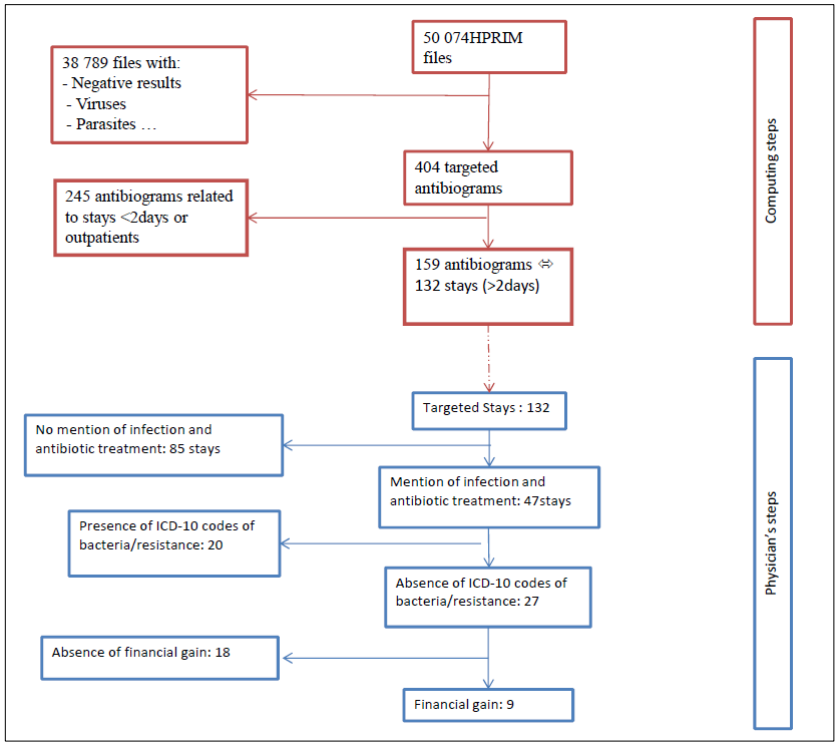


Figure 2. Study steps flow chart (Red: Antimicrobial susceptibility tests selecting algorithm; Blue: Description of checking and improving coding steps)

4. Discussion

With a sensitivity of 42.5%, coding bacteria and resistance remains incomplete, which is consistent with a study carried on the 2014 French case-mix database that showed nationwide inconsistencies between the type of infections coded and the associated bacteria and resistance. Although only 9 out of 132 stays have been upgraded, the recovered benefit was important, corresponding for instance to the total employer cost of a nurse or a technician over one year.

Our results show a high specificity of initial coding in the PMSI database since only 2 stays were incorrectly coded, whose deletion of the corresponding codes led to no financial loss. In addition to searching for missing codes, we also aim to check codes already present. Indeed, removing incorrect codes allows avoiding financial penalties from the national Health Insurance resulting of controls of the consistency between diagnoses in the case-mix database, and information available on the medical records.

Usually, predicting PMSI coding exploits supervised machine learning methods over unstructured data like medical observation. Nevertheless, Scheurwegs et al. used structured data such as biological test results and drug prescription to improve the machine learning results [5]. However, machine learning methods identify data co-occurrences, which could lead to association rules that do not necessarily reflect expert

knowledge. The decision rules implemented in our approach were defined from the regulatory ones established in France by the national Health Insurance, which makes them intuitive to understand.

To our knowledge, this is the first time that a proof of concept is provided for supporting the efficiency and the return on investment of implementing automated analysis of available electronic data to improve coding of bacterial infection in a case-mix database. Several works that involved different data sources such as microbiology, pharmacy and case mix database have been published [6,7] about the surveillance of healthcare associated infections. However, emphasis was on surveillance, and not on improvement of coding. Clinical decision support systems have previously supported antimicrobial stewardship programs by collating data from multiple sources, including microbiology and pharmacy [8]. Part of the processing of these systems is based on the analysis of new culture results when they become available, which results in the generation of alerts regarding resistant pathogens. However, it differs from our approach as it intends to improve both antimicrobial prescription and appropriateness. As for electronic surveillance systems, it was acknowledged that efficient processing was dependent on their ability to interface with institutional data sources.

We propose to work directly on messages transmitted by the LMS containing the antimicrobial susceptibility tests rather than extracting this information from a database. This strategy requires a complete interoperability between the LMS and the coding tool. In our hospital, such messages are based on the HPRIM format. This format is not a standard at the international level but is close to HL7 V2.x that is implemented in many hospital information systems worldwide. For this reason, we are confident that a similar approach could be implemented in other clinical health information systems, even if they do not use the HPRIM protocol.

References

[1] Gedeborg R, Furebring M, Michaëlsson K. Diagnosis-dependent misclassification of infections using administrative data variably affected incidence and mortality estimates in ICU patients. J Clin Epidemiol. 2007;60(2):155.e1-e155.e11. doi:10.1016/j.jclinepi.2006.05.013

[2] Pinyo Rattanaumpawan MD, Wongkamhla T, Thamlikitkul V. Accuracy of ICD-10 coding system for identifying comorbidities and infectious conditions using data from a Thai University Hospital administrative database. J Med Assoc Thai. 2016;99(4):368-373.

[3] Wilhelms SB, Walther SM, Huss F, Sjöberg F. Severe sepsis in the ICU is often missing in hospital discharge codes. Acta Anaesthesiol Scand. 2017;61(2):186-193. doi:10.1111/aas.12814

[4] Chan M, Lim PL, Chow A, Win MK, Barkham TM. Surveillance for Clostridium difficile Infection: ICD-9 Coding Has Poor Sensitivity Compared to Laboratory Diagnosis in Hospital Patients, Singapore. Braga EM, ed. PLoS ONE. 2011;6(1):e15603. doi:10.1371/journal.pone.0015603

[5] Scheurwegs E, Luyckx K, Luyten L, Daelemans W, Van den Bulcke T. Data integration of structured and unstructured sources for assigning clinical codes to patient stays. J Am Med Inform Assoc. 2016;23(e1):e11-e19. doi:10.1093/jamia/ocv115

[6] de Bruin JS, Seeling W, Schuh C. Data use and effectiveness in electronic surveillance of healthcare associated infections in the 21st century: a systematic review. J Am Med Inform Assoc. 2014;21(5):942-951. doi:10.1136/amiajnl-2013-002089

[7] Freeman R, Moore LSP, García Álvarez L, Charlett A, Holmes A. Advances in electronic surveillance for healthcare-associated infections in the 21st Century: a systematic review. J Hosp Infect. 2013;84(2):106-119. doi:10.1016/j.jhin.2012.11.031

[8] Forrest GN, Van Schooneveld TC, Kullar R, Schulz LT, Duong P, Postelnick M. Use of Electronic Health Records and Clinical Decision Support Systems for Antimicrobial Stewardship. Clin Infect Dis. 2014;59(suppl_3):S122-S133. doi:10.1093/cid/ciu565

Challenges of Trustable AI and Added-Value on Health
B. Séroussi et al. (Eds.)

doi:10.3233/SHTI220448

Artificial Intelligence in Kidney Transplantation: A Scoping Review

Asma ALAMGIR [a], Hagar HUSSEIN [a], Yasmin ABDELAAL [a], Alaa ABD-ALRAZAQ [a] and Mowafa HOUSEH [a,1]
[a] *College of Science and Engineering, Hamad Bin Khalifa University*

Abstract: Artificial Intelligence (AI) technologies are increasingly being used to enhance kidney transplant outcomes. In this review, we explore the use of AI in kidney transplantation (KT) in the existing literature. Four databases were searched to identify a total of 33 eligible studies. AI technologies were used to help in diagnostic, predictive and medication management purposes for kidney transplant patients. AI is an emerging tool in KT, however, there is a research gap exploring the limitations associated with implementing AI technologies in the field. Research is also needed to recognize clinical educational needs and other barriers to promote adoption and standardization of care for KT patients amongst clinicians.

Keywords: Artificial Intelligence, Machine Learning, Deep Learning, Kidney Transplant.

1. Introduction

About 9.1% of the world's population suffer from chronic kidney failure requiring dialysis or a kidney transplant [1]. Kidney transplantation (KT) is the preferred intervention as it allows patients a better quality of life while being cost effective compared to long-term dialysis [2]. The development and use of artificial intelligence (AI) in medicine, and clinical applications is evolving in several fields, including the field of kidney disease and KT [3-6]. AI appears to be a promising tool in healthcare and clinical decision support, providing personalized diagnostics, therapeutic solutions, and predictions of future events such as hospitalization and patient's survival [3].

A review of collective evidence exploring the use of AI in KT is limited. We came across two reviews; Burlacu et al. (2020), reviews AI in nephrology, including in KT [6]. The search for the review was conducted in August 2019, omitting evidence related to KT during the pandemic. The review also does not report the types of AI technologies and algorithms observed. Seyahi et al. (2021) also explores AI used in KT [7]. The authors use only one data source to conduct their study and solely address AI applications. Our scoping review cover these gaps while combining the latest evidence to help keep clinicians informed and recognize future research opportunities.

The aim of this scoping review is to explore: 1) How is AI being used in KT? and, 2) What are the characteristics of AI technologies utilized for kidney transplant purposes?

[1] Corresponding author, Mowafa HOUSEH, College of Science and Engineering, Hamad Bin Khalifa University, Doha-Qatar, Email: mhouseh@hbku.edu.qa

2. Methods

The scoping review is in line with PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) Extension for Scoping Reviews guideline [8]. We searched four databases between 1st till 5th March 2021. While manually checking citations on Google Scholar, only the first 40 citations were relevant to the subject of interest. Hence, the top 40 hits are included. The search engine is known to retrieve many hundreds of citations in order of relevance to the search terms. This method can be seen in another scoping review [9]. Table 1 displays details of the search strategy used.

Only primary studies conducted in English between the years 2018 till 2021 were included in this scoping review. Three reviewers independently screened the articles using Rayyan System Inc. In case of conflict, discussion was held among reviewers to come to a mutual consensus. The reviewers independently extracted data using a data extraction form. The findings are synthesized narratively, then classified and described in terms of their purpose and characteristics.

3. Results

Only 32 studies met the criteria and are included in this review. One additional study was added by backward citation of the reference list of the included studies. Most studies were published between 2020-2019 (n=28) and conducted in the United State of America (n =11). Table 2 displays the common AI branch and models observed in the studies.

The use of AI technologies for KT found in included studies can be categorized into three – diagnoses, prediction, and prescription.

3.1 AI to Diagnose Kidney Transplant Patients

A total of 18 studies investigated usefulness of AI models for diagnostic purposes in kidney transplant patients [10,11,20-27,12-19]. For example, Kanzelmeyer et al. (2019) investigated use of AI in early diagnoses of chronic active antibody mediated rejection (cABMR) in kidney transplant patient [23]. Another example, study by Shehata et al. (2019) used AI to distinguish between diagnoses of non-rejection and acute rejection in transplant patients [19]. Majority of the models were based on RF [20,21,24,25] and SVM (Support Vector Machine) [18,21,23] models. Most data sources used to train and test the AI models were based on clinical setting such as hospitals (n =12). Clinical data such as blood, and biopsies (n =13) were the most used data type.

3.2 AI as a Prediction Tool in Kidney Transplantation

A total of 12 studies used AI as a prediction tool [28,29,38,39,30-37]. Algorithms and datasets were used to predict survival of graft tissue, rejection, and delayed graft function. Studies also explored AI facilitated prediction for critical clinical decisions of weighing benefits of receiving an available kidney over waiting for a 'better offer'. Information of patients and donors are used to predict suitable match and weigh risks of undergoing the procedure or receiving a certain quality of kidney.

Table 1. Search Strategy

Database/ # Citations	Search Strategy
Embase 349 Citations	('artificial intelligence'/exp OR 'machine learning'/exp OR 'deep learning'/exp OR 'neural network*' OR 'supervised learning'/exp OR 'unsupervised learning'/exp OR 'natural language' OR 'data mining'/exp) AND ('kidney transplant*':ti,ab,kw OR 'renal transplant*':ti,ab,kw OR 'kidney graft*':ti,ab,kw OR 'kidney allograft':ti,ab,kw)
CINAHL 25 Citations	('kidney transplant*' OR 'renal transplant*' OR 'kidney graft*' OR 'kidney allograft*') AND ('artificial intelligence' OR 'machine learning' OR 'deep learning' OR 'neural network*' OR 'supervised learning' OR 'unsupervised learning' OR 'natural language' OR 'data mining') 2018 – 2021
Pubmed 91 Citations	("Artificial intelligence"[Title/Abstract] OR "Machine learning"[Title/Abstract] OR "Deep learning"[Title/Abstract] OR "Neural network*"[Title/Abstract] OR "Supervised learning"[Title/Abstract] OR "Unsupervised learning"[Title/Abstract] OR "Natural language processing"[Title/Abstract] OR "Data mining"[Title/Abstract]) AND ("Kidney Transplant*"[Title/Abstract] OR "Renal Transplant*"[Title/Abstract] OR "Kidney Graft*"[Title/Abstract] OR "Kidney Allograft*"[Title/Abstract])
Google Scholar 40 Citations	(“Kidney Transplant*” OR “Renal Transplant*” OR “Kidney Graft* OR “Kidney Allograft*”) AND ("Artificial intelligence" OR "Machine learning" OR "Deep learning" OR "Data Mining")

Table 2. Most common features of AI-based techniques used for kidney transplantation

Features	Studies (N=33)
AI branch [a]	
Deep Learning	8
Machine Learning	25
Natural Language Processing	1
AI models / algorithm [b]	
Random Forest	11
Logistic Regression	6
Gradient Boosting	4
Support Vector Machine	4
Artificial Neural Network	3

[a] Numbers do not add up as some studies were based on more than one AI branch
[b] Numbers display only the most common models/algorithm used.

3.3 AI used as a Prescription Tool in Kidney Transplantation

AI techniques were used to manage appropriate dosage of immunosuppressants and other medications [40-42]. The models used included fuzzy logic [42], RF [40], and ANN [41]. All studies used clinical setting as the form of data to test and train their models. Genetic data was used in 2 studies. AI was used to manage immunosuppressant dosage to improve efficiency in KT patients. One study used a different approach by using ANN to understand the relationship between genetic factors and tacrolimus dose [41].

5. Discussion

With the increasing adaption of electronic health record, there is an abundance of patient data providing AI technologies the platform to improve multiple aspects of KT care. Clinicians and their input are a big part of integrating any system to enhance the health experience of patients, however, none of our included studies studied the clinician’s perspective of using AI or the challenges and the ethics that need to be considered. Moreover, clinicians may not fully understand the technical explanations of these

algorithms or performance metrices. Training programs and other resources catered towards healthcare providers are necessary. AI could be the answer to the shortage of solid organs as it provides the answer to many of the complications and challenges attached to it. Therefore, proactiveness to include stakeholders is needed, while also exploring barriers and facilitators of integrating AI technologies in the clinical setting.

5. Conclusion

Kidney transplantation is a complex intervention requiring patients and clinicians to undergo multiple processes and critical decision making. AI is proving to be an important tool to support clinicians and patients to make the best decision for their needs. AI is being used for diagnoses, making predictions, and prescribing personalized care plans for KT patients. More research is required to promote AI adoption within the field.

References

[1] Carney EF. The impact of chronic kidney disease on global health. Nat Rev Nephrol 2020 165 [Internet]. 2020 Mar 6;16(5):251–251.
[2] Klarenbach S, Manns B. Economic evaluation of dialysis therapies. Semin Nephrol [Internet]. 2009 Sep [cited 2022 Jan 21];29(5):524–32. Available from: https://pubmed.ncbi.nlm.nih.gov/19751898/
[3] Chaudhuri S, Long A, Zhang H, Monaghan C, Larkin JW, Kotanko P, et al. AI enabled applications in kidney disease. Semin Dial. 2021;34(1):5–16.
[4] Hannun PGC, Andrade LGM. The future is coming: promising perspectives regarding the use of machine learning in renal transplantation. J Bras Nefrol. 2019;41(2):284–7.
[5] Becker JU, Mayerich D, Padmanabhan M, Barratt J, Ernst A, Boor P, et al. Artificial intelligence and machine learning in nephropathology. Kidney Int. 2020;98(1):65–75.
[6] Burlacu A, Iftene A, Jugrin D, Popa IV, Lupu PM, Vlad C, et al. Using AI Resources in Dialysis and KT Patients: A Literature Review. Vol. 2020, BioMed Research International. Hindawi Limited; 2020.
[7] Seyahi N, Ozcan SG. AI and KT. World J Transplant. 2021 Jul 1, 11(7).
[8] Tricco AC, Lillie E, Zarin W, O'Brien KK, Colquhoun H, Levac D, et al. PRISMA Extension for Scoping Reviews (PRISMA-ScR): Checklist and Explanation. Ann Intern Med. 2018 Sep;169(7):467–73.
[9] Abd-alrazaq AA, Alajlani M, Alalwan AA, Bewick BM, Gardner P, Househ M. An overview of the features of chatbots in mental health: A scoping review. Int J Med Inform. 2019 Dec 1;132:103978.
[10] Zubke M, Katzensteiner M, Bott OJ. Integration of Unstructured Data into a Clinical Data Warehouse for Kidney Transplant Screening - Challenges & Solutions. Stud Heal Technol Inf. 2020;270:272–6.
[11] Shehata M, Shalaby A, Switala AE, El-Baz M, Ghazal M, et al. A multimodal computer-aided diagnostic system for precise identification of renal allograft rejection: Preliminary results. Med Phys. 2020;47(6).
[12] Vittoraki AG, Fylaktou A, Tarassi K, Tsinaris Z, Petasis GC, Gerogiannis D, et al. Patterns of 1,748 Unique Human Alloimmune Responses Seen by Simple ML Algorithms. Front Immunol. 2020;11.
[13] Costa SD, Andrade LGM, Barroso FVC, Oliveira CMC, Daher EDF, Fernandes PFCBC, et al. The impact of deceased donor maintenance on delayed kidney allograft function: A ML analysis. PLoS One. 2020;15(2).
[14] Abdeltawab H, Shehata M, Shalaby A, Khalifa F, Mahmoud A, El-Ghar MA, et al. A Novel CNN-Based CAD System for Early Assessment of Transplanted Kidney Dysfunction. Sci Rep. 2019;9(1):5948.
[15] Zacharias HU, Altenbuchinger M, Schultheiss UT, Samol C, Kotsis F, Poguntke I, et al. A Novel Metabolic Signature To Predict the Requirement of Dialysis or Renal Transplantation in Patients with Chronic Kidney Disease. J Proteome Res. 2019 Apr;18(4):1796–805.
[16] Zou Y, Zhang W, Zhou HH, Liu R. Analysis of long noncoding RNAs for acute rejection and graft outcome in kidney transplant biopsies. Biomarkers Med. 2019;13(3):185–95.
[17] Aubert O, Higgins S, Bouatou Y, Yoo D, Raynaud M, Viglietti D, et al. Archetype Analysis Identifies Distinct Profiles in Renal Transplant Recipients with Transplant Glomerulopathy Associated with Allograft Survival. J Am Soc Nephrol. 2019 Apr;30(4):625–39.
[18] Wittenbrink N, Herrmann S, Blazquez-Navarro A, Bauer C, Lindberg E, Wolk K, et al. A novel approach reveals that HLA class 1 single antigen bead-signatures provide a means of high-accuracy pre-transplant

risk assessment of acute cellular rejection in renal transplantation. BMC Immunol. 2019 Apr;20(1):11.
[19] Shehata M, Khalifa F, Soliman A, Ghazal M, Taher F, El-Ghar MA, et al. Computer-Aided Diagnostic System for Early Detection of Acute Renal Transplant Rejection Using Diffusion-Weighted MRI. IEEE Trans Biomed Eng. 2019;66(2):539–52.
[20] Friedewald JJ, Kurian SM, Heilman RL, Whisenant TC, Poggio ED, Marsh C, et al. Development and clinical validity of a novel blood-based molecular biomarker for subclinical acute rejection following kidney transplant. Am J Transplant. 2019;19(1):98–109.
[21] Reeve J, Böhmig GA, Eskandary F, Einecke G, Gupta G, Madill-Thomsen K, et al. Generating automated kidney transplant biopsy reports combining molecular measurements with ensembles of machine learning classifiers. Am J Transplant. 2019;19(10):2719–31.
[22] Archdekin B, Sharma A, Gibson IW, Rush D, Wishart DS, Blydt-Hansen TD. Non-invasive differentiation of non-rejection kidney injury from acute rejection in pediatric RT recipients. Pediatr Transplant. 2019;23.
[23] Kanzelmeyer NK, Zürbig P, Mischak H, Metzger J, Fichtner A, Ruszai KH, et al. Urinary proteomics to diagnose chronic active antibody-mediated rejection in pediatric KT. Transpl Int. 2019;32(1).
[24] Bae S, Massie AB, Thomas AG, Bahn G, Luo X, Jackson KR, et al. Who can tolerate a marginal kidney? Predicting survival after deceased donor kidney transplant by donor-recipient combination. Am J Transplant Off J Am Soc Transplant Am Soc Transpl Surg. 2019 Feb;19(2):425–33.
[25] Shaikhina T, Lowe D, Daga S, Briggs D, Higgins R, Khovanova N. Decision tree and RF models for outcome prediction in antibody incompatible KT. Biomed Signal Process Control. 2019;52:456–62.
[26] Rashidi P, Bagherzadeh J, Niazkhani Z, Pirnejad H. A dynamic model for predicting graft function in kidney recipients' upcoming follow up visits: A clinical application of ANN. Int J Med Inform. 2018;119:125–33.
[27] Marsh JN, Matlock MK, Kudose S, Liu T-C, Stappenbeck TS, Gaut JP, et al. DL global glomerulosclerosis in transplant kidney frozen sections. IEEE Trans Med Imaging. 2018;37(12).
[28] Kim S-K, Yang JJ, Hwang S-H, Sung H, Shin S, Ko S-Y, et al. Prediction of HLA-DQ in Deceased Donors and its Clinical Significance in Kidney Transplantation. Ann Lab Med. 2021 Mar;41(2):190–7.
[29] Wang D, Jing Z, He K, Garmire LX. Cox-nnet v2.0: improved neural-network based survival prediction extended to large-scale EMR data. Bioinformatics. 2021 Jan;
[30] Placona AM, Martinez C, McGehee H, Carrico B, Klassen DK, Stewart D. Can donor narratives yield insights? A NL processing proof of concept to facilitate kidney allocation. Am J Transplant. 2020;20(4).
[31] Luo Y, Tang Z, Hu X, Lu S, Miao B, Hong S, et al. Machine learning for the prediction of severe pneumonia during posttransplant hospitalization in recipients of a deceased-donor kidney transplant. Ann Transl Med Vol 8, No 4 (February 2020) Ann Transl Med (Focus "Intrauterine Adhes. 2020;
[32] Bae S, Massie AB, Caffo BS, Jackson KR, Segev DL. Machine learning to predict transplant outcomes: helpful or hype? A national cohort study. 2020;
[33] Kawakita S, Beaumont JL, Jucaud V, Everly MJ. Personalized prediction of delayed graft function for recipients of deceased donor kidney transplants with machine learning. Sci Rep. 2020 Oct;10(1):18409.
[34] Peng B, Gong H, Tian H, Zhuang Q, Li J, Cheng K, et al. The study of the association between immune monitoring and pneumonia in kidney transplant recipients through machine learning models. J Transl Med. 2020;18(1):N.PAG-N.PAG.
[35] Massie AB, Boyarsky BJ, Werbel WA, Bae S, Chow EKH, Avery RK, et al. Identifying scenarios of benefit or harm from KT during the COVID-19 pandemic: A stochastic simulation and machine learning study. Am J Transplant Off J Am Soc Transplant Am Soc Transpl Surg. 2020 Nov;20(11).
[36] Elihimas UF, Couto JP, Pereira W, Barros de Oliveira Sá MP, Tenório de França EE, Aguiar FC, et al. LR Model in a ML Application to Predict Elderly Kidney Transplant Recipients with Worse Renal Function One Year after KT: Elderly KTbot. Grosset J-F, editor. J Aging Res. 2020;2020:7413616.
[37] Atallah D, Badawy M, El-Sayed A. A new proposed feature selection method to predict kidney transplantation outcome. Health Technol (Berl). 2019 Nov;9:847–856.
[38] Rashidi P, Bagherzadeh M J, Niazkhani Z, Pirnejad H. Predicting the function of transplanted kidney in long-term care processes:Application of a hybrid model. J Biomed Inform. 2019 Mar;91:103116.
[39] Mark E, Goldsman D, Gurbaxani B, Keskinocak P, Sokol J. Using machine learning and an ensemble of methods to predict kidney transplant survival. PLoS One. 2019 Jan;14(1):e0209068.
[40] Cai N, Zhang X, Zheng C, Zhu L, Zhu M, Cheng Z, et al. A novel random forest integrative approach based on endogenous CYP3A4 phenotype for predicting tacrolimus concentrations and dosages in Chinese renal transplant patients. J Clin Pharm Ther. 2020;45(2):318–23.
[41] Thishya K, Vattam KK, Naushad SM, Raju SB, Kutala VK. Artificial neural network model for predicting the bioavailability of tacrolimus in patients with renal transplantation. PLoS One. 2018;13(4).
[42] Improta G, Mazzella V, Vecchione D, Santini S, Triassi M. Fuzzy logic–based clinical decision support system for the evaluation of renal function in post-transplant Patients. J Eval Clin Pr. 2020;26(4):1224.

Challenges of Trustable AI and Added-Value on Health
B. Séroussi et al. (Eds.)

doi:10.3233/SHTI220449

Use of e-Health in Norwegian FACT Teams: A User Perspective

Erlend BØNES [a, b,1], Conceição GRANJA [a, b] and Terje SOLVOLL [a, b]
[a] *University Hospital of North Norway, Norwegian Centre for e-health Research, Tromsø, Norway*
[b] *Nord University, Faculty of Nursing and Health Sciences, Bodø, Norway*

Abstract. Flexible assertive community treatment (FACT) is a model for delivering long-term, integrated and comprehensive treatment and follow-up for patients with severe mental illness. The objective of this study was to examine ICT challenges of Norwegian FACT teams. Doing observations in 3 teams and interviews with 5 teams we examined use of ICT systems, identifying challenges with the use of the electronic whiteboards, electronic health records, and team calendars. Better ICT systems and infrastructure are needed to support Norwegian FACT teams.

Keywords. Mental health, FACT, electronic health records, electronic whiteboards, eHealth, video consultations

1. Introduction

Flexible assertive community treatment (FACT) is a model for delivering long-term, integrated and comprehensive treatment and follow-up for patients with severe mental illness [1]. This model has been implemented in the Netherlands, Norway, Sweden, England and Denmark [2]. FACT teams are multidisciplinary and should consist of a psychiatrist, a psychologist, case managers, an employment specialist, an addiction specialist and a peer support worker [3]. Norwegian FACT teams often include members employed in both primary care and specialist care. They include a team leader, and team coordinator. The team coordinator is responsible for the overall administrative work in the teams.

Patients at risk of relapse or readmission receive intensive follow-up from FACT teams, while patients who are in more stable condition receive case management. The patients who receive intensive follow-up are discussed in daily meetings, where the team discuss the status of the patient, and plan further treatment. During these meetings, an electronic whiteboard is used to keep track of patients. Access to relevant patient information from electronic health records (EHRs) is essential for health care workers. In Norway, specialist health care in three of the four health care regions use the EHR system, DIPS AS, while the fourth are using DocuLive provided by Siemens AS. Within primary care there are several different EHR systems in use.

In 2020, there were approximately 70 FACT teams in Norway [2]. However, an evaluation of Norwegian FACT teams [4], shows that there are some issues for the teams

[1] Corresponding Author, Erlend Bønes, Norwegian Centre for e-Health Research, University Hospital of North Norway, PO Box 35, Tromsø, Norway; E-mail: erlend.bones@ehealthresearch.no.

using information and communication technology (ICT) solutions. This paper reports on the challenges of Norwegian FACT teams using ICT solutions.

2. Methods

Five Norwegian FACT teams were included in the project. The teams were selected based on purposeful selection; a strategy intended to get relevant information related to the research goals [5]. To ensure representativeness of the teams, we selected two urban teams and three rural teams with representative number of patients. The selected teams were in different geographical areas of Norway.

Using a Computer-Supported Cooperative Work (CSCW) framework, one of the authors did observations and semi-structured interviews in 3 of the teams, following an ethnographic approach. This is a well-established approach for the application of the CSCW framework in the health care field [6-8]. Due to the restrictions related to the covid-19 pandemic, we were not able to do observations in two of the teams. In these two teams we only did semi-structured interviews using Skype for business. The observations and interviews were completed between August 2020 and January 2021.

During the physical observations, one of the authors followed the teams for 3 to 5 workdays each. The author participated in the daily whiteboard meetings and other team meetings and observed use of the ICT solutions. He observed meetings with one patient in each team. The patients participating were selected by the teams. Only one patient consented for each team. The author had informal conversations with the team members. During the observations, the author wrote memos of what he observed and added his own ideas about use of ICT tools. The semi-structured interviews were done with the team leaders and team coordinators in the five selected teams. Initial topics for the interview was based on previous knowledge of the ICT implementation of FACT teams in Norway [9]. This included Electronic whiteboards, EHR systems, Video conferencing systems, and Mobile devices. During the interview with Team 1, one additional topic was identified and included in the later interviews. Notes were taken during the interview and expanded on immediately after the interviews. Data from the observations and interviews were analyzed in several stages during and after the data collection. This process identified preliminary themes of data, used to categorize the data.

3. Results

Table 1 shows the characteristics of the different teams and data collection methods.

Table 1. characteristics of the teams and methods of data collection

Team	Urban/rural	Coverage area	Team employment	Methods used
1	Rural	4 municipalities	Specialist and secondary care.	Observations and interviews
2	Urban	One city district	Specialist and secondary care.	Observations and interviews
3	Urban	One municipality, including one town	Specialist and secondary care.	Observations and interviews

4	Rural	2 municipalities	Specialist and secondary care.	Video interviews
5	Rural	2 municipalities	Specialist care only.	Video interviews

During the observations, the author observed use of the electronic whiteboards, EHR systems, calendar systems and video conferencing solutions. In the interviews, the team leaders and team coordinators were asked to describe the use of different types of ICT solutions in daily practice, based on the identified topics. During the interview with Team 1, team calendars was identified as as an additional type of ICT solution relevant for FACT teams. This topic was added in the interviews with the remaining teams. These questions were used as a starting point to discuss strengths and weaknesses of the ICT solutions. The interviewer asked about the biggest challenges related to use of ICT, and how ICT could be used to make the teams work easier. Additionally, the interviewees were asked to describe further functional needs currently not available in the current ICT solutions, and how they communicate with their patients. The themes of data used in the pre-analysis of the data were same as the ICT solutions: Electronic whiteboards, EHR systems, Video conferencing systems, Team calendar and Mobile devices.

3.1. Electronic whiteboards

The teams that were a part of the study used different electronic whiteboard solutions; they were all based on Microsoft Excel. Teams found the electronic whiteboards hard to use. One informant said that "you do not recognise the FACT model" when looking at the board. They wanted the whiteboard to represent the model in a better way, for instance by showing more information about standardized patient pathways. Team 1 and 3 reported that when a patient is not in need of intensive follow-up, and is moved to case management, some information stored about the patient will be lost. All teams described the whiteboard solutions in use as stand-alone systems without any integration to the EHRs or other ICT solutions used by the teams. At the same time, all teams reported a need for the electronic whiteboard to be integrated to the EHRs.

3.2. Electronic health records

Team 5 only had employees in specialist care and consequently only documented in DIPS. The other four teams had employees from specialist and primary care. Team 2 only documented in DIPS. The remaining three teams documented in both DIPS and EHR systems for primary care. Team 1 documented in more than one EHR system for primary care since it covered four municipalities. To give FACT team members who work primary care access, they often had so called zero percent positions or simplified employment in specialist care to be given organisational access to DIPS. In addition to this, in most teams they use a virtual private network (VPN) solution like Citrix to get technical access. The teams that use Citrix to get access to DIPS, found this an annoyance, especially if they were using a mobile device. Team 2 were worried that only documenting in DIPS would lead to their work being hidden from the municipalities, and as a worst case that this may lead to a halt in funding. The rural teams with long travel distances often postpone documenting a patient meeting in the EHR to the next day. Teams 1 pointed out cultural differences between team members in primary care and specialist care, with members from specialist care more used to extensive documentation in the EHR.

3.3. Video consultations

Teams 1, 3 and 4 used video consultations to some of their patients. The technical solutions used depended on what was recommended for clinical use in their regions, Whereby in the North region and Pexip in the South East region. The teams that used video conference emphasised that it is not suitable for all patients. Also, Team 3 told us that not all patients have access to their own equipment. In most cases, the use of video conference was planned, but Team 1 described a situation where video conference was used in an acute situation, to establish a video consultation between a team member who was with a patient to the team psychiatrist who was in another municipality. Team 2 did not use video consultations and said that their patients do not wish to use it due to mental health issues and a lack of equipment. Team 5 also did not use video consultations, and said they doubted it is suited for their patient group, and that many patients lack phones or Internet access.

3.4. Team calendar

Team members use many different calendars, including the calendar in DIPS, and various versions on Microsoft Outlook depending on their employment. Team 2 used physical calendar books, and said they are completely dependent on them. Team 5 had DIPS on a laptop PC, where they could see the calendars, but since they need an internet connection for access, it is not always available when traveling. Some teams reported a need to have an overview of where the team members are. One reason for this was to be able to coordinate the team better. Another reason was that it improves safety because it gives an overview of when team members visit unstable patients. One team said that they already had a good overview of when and team members visit unstable patients, so they did not see a need for a such a system.

3.5. Mobile devices

Some of the teams have laptop PCs that they can bring when they are visiting patients. Team 1 reported that it was something that they wanted but did not have yet. Team 3 bring the laptop with them sometimes when visiting patients, but usually not, because they usually do not have network access when visiting patients. They said that Internet-connected PC could be useful for helping the patients with practical things like online banking and social services web sites.

4. Discussion

Our study has uncovered various issues with ICT solutions for Norwegian FACT teams. FACT teams do not think current whiteboard solutions fully meet their needs. The main issue with the electronic whiteboards is that there is no integration towards the existing EHR systems. If such an integration was in place, it could allow relevant data to be transferred automatically. This could include documentation of patient treatment, patient ID's and healthcare worker ID's. Whiteboards should also include information about standardized patient pathways. For many Norwegian FACT teams there are several different EHR systems that contain relevant information about their patients. There is a

need for systems where FACT team members have easy access to relevant information from EHR systems. The same data does not need to be stored in several different systems but should be able to be displayed in the relevant systems. Since FACT team members are highly mobile, patient data should be available from a mobile device, like phones, laptop computers or tablets. While Norwegian laws allow sharing of relevant data, the e-health infrastructure that is in place does not support such sharing [9]. This means that EHR data is usually easily accessible for healthcare workers in the institution that is responsible for the data, but harder to access for other healthcare workers. To circumvent this, healthcare workers who need access to data from an institution they do not work in, have sometimes been hired in so called zero percent positions in the institution, as a workaround. Some FACT teams use video consultations when treating patients. Video conference is not suited for all patients in FACT teams, or in all situations. Some patients also lack the access to equipment needed, but it is a useful tool in some situations. Currently, there are several different calendar solutions FACT teams needs to use. Calendar information should be displayed in one easy to access system. This would make coordination within the team easier and provide additional safety for team members. Having access to a mobile device like a laptop or PC is helpful in many situations when visiting patients, both for accessing the EHR, and practical help like online banking and social services. However, this requires Internet connection for the device, which may be a challenge when travelling.

5. Conclusion

FACT teams have been successfully implemented in Norway, despite that FACT team members see issues and challenges with their ICT solutions. Common requirements from the FACT teams are integration of electronic whiteboards and EHR systems, easy access to relevant EHR information and better calendar solutions. This paper only reports on a preliminary analysis of the data. The data reported in this paper needs to be analysed in-depth to inform the definition of the requirements for the use of e-health in FACT teams.

References

[1] van Veldhuizen JR. FACT: a Dutch version of ACT, Community Ment Health J 43 (2007), 421-433.
[2] Trane K, Aasbrenn K, Rønningen M, et al. Flexible assertive community treatment teams can change complex and fragmented service systems: experiences of service providers, International journal of mental health systems 15 (2021), 38.
[3] van Veldhuizen JR, FACT - Visjon, modell og organisering av FACT-modellen, in, Nasjonal kompetansetjeneste for samtidig rusmisbruk og psykisk lidelse, Brumunddal 2013.
[4] Landheim A, Odden S. Evaluering av FACT-team i Norge – Sluttrapport in, Nasjonal kompetansetjeneste for samtidig rusmisbruk og psykisk lidelse (NKROP), Hamar, 2020.
[5] Maxwell J. Qualitative Research Design: An Interactive Approach, Sage, Los Angeles, 2013.
[6] Blomberg J, Karasti H. Reflections on 25 Years of Ethnography in CSCW, Computer Supported Cooperative Work (CSCW) 22 (2013), 373-423.
[7] Pratt W, Reddy MC, McDonald DW, et al. Incorporating ideas from computer-supported cooperative work, Journal of Biomedical Informatics 37 (2004), 128-137.
[8] Fitzpatrick G, Ellingsen G. A review of 25 years of CSCW research in healthcare: contributions, challenges and future agendas, Computer Supported Cooperative Work (CSCW) 22 (2013), 609-665.
[9] Bønes E, Granja C, Solvoll T. Implementation of the Flexible Assertive Community Treatment (FACT) Model in Norway: eHealth Assessment Study, Journal of medical Internet research 24 (2022), e32220.

Challenges of Trustable AI and Added-Value on Health
B. Séroussi et al. (Eds.)

doi:10.3233/SHTI220450

Determination of Cut Off for Endometrial Thickness in Couples with Unexplained Infertility: Trustable AI

Ameneh MEHRJERD[a,b], Hassan REZAEI[b], Saeid ESLAMI[a,c,d,1] and Nayyere KHADEM GHAEBI[a]

[a] *Department of Medical Informatics, Faculty of Medicine, Mashhad University of Medical Sciences, Mashhad, Iran*

[b] *Department of Computer Science, Faculty of Mathematics, Statistics and Computer Science, University of Sistan and Baluchestan, Zahedan, Iran*

[c] *Associate Professor at Department of Medical Informatics, Faculty of Medicine, Mashhad University of Medical Sciences, Mashhad, Iran*

[d] *Full Professor at Department of Obstetrics & Gynecology, School of Medicine, Mashhad University of Medical Sciences, Mashhad, Iran*

Abstract. Endometrial thickness in assisted reproductive techniques is one of the essential factors in the success of pregnancy. Despite extensive studies on endometrial thickness prediction, research is still needed. We aimed to analyze the impact of endometrial thickness on the ongoing pregnancy rate in couples with unexplained infertility. A total of 729 couples with unexplained infertility were included in this study. A random forest model (RFM) and logistic regression (LRM) were used to predict pregnancy. Evaluation of the performance of RFM and LRM was based on classification criteria and ROC curve, Odd Ratio for ongoing Pregnancy by EMT categorized. The results showed that RFM outperformed the LRM in IVF/ICSI and IUI treatments, obtaining the highest accuracy. We obtained a 7.7mm cut-off point for IUI and 9.99 mm for IVF/ICSI treatment. The results showed machine learning is a valuable tool in predicting ongoing pregnancy and is trustable via multicenter data for two treatments. In addition, Endometrial thickness was not statistically significantly different from CPR and FHR in both treatments.

Keywords. Machine learning, Endometrial thickness, Unexplained, Prediction model

1. Introduction

Infertility was considered after the second half of the 19th century B. C., and various causes and treatments were gradually introduced for it [2]. The cause of infertility, such as male and female factors, unexplained infertility, genetic factors, or a combination of these factors, are considered. Unexplained infertility is a factor in which standard tests such as tubing and ovulation and sperm testing are standard for a couple but still infertile

[1] Corresponding Author, Saeid Eslami, Department of Medical Informatics, Faculty of Medicine, Mashhad University of Medical Sciences, Mashhad, Iran; E-mail: s.eslami.h@gmail.com/ EslamiS@mums.ac.ir

[11]. Unexplained factor makes up about 15 to 30% of infertile couples [1]. Endometrial thickness (EMT) factor is also one of the essential factors in influencing pregnancy rates for treatments that have been evaluated in various articles [3; 4; 6; 7; 12]. For example, in study, Quaas et al. (2021) evaluated the effect of EMT on pregnancy under ovarian stimulation protocol used in IUI treatment for couples with unexplained infertility [10]. Herein, we examined EMT and its effect on pregnancy in IUI and IVF/ ICSI treatments. Also, our study applied a machine learning perspective to construct a superior model to evaluate the impact of EMT on pregnancy rate.

2. Materials and Methods

2.1. Patient selection

Couples treated from two infertility centers were included in this study. Infertility treatments using sperm or eggs donated and surrogate uterus were also excluded from the study. A total of 729 couples (952 treatment cycles) with unexplained infertility, include 641 IUI cycles (500 couples) with unexplained infertility and 311 IVF/ICSI cycles (229 couples). Only the first three treatment cycles were considered. Some of data for IVF/ICSI and IUI can be seen in Table 1 and Table 2, respectively. This study was approved by the Institutional Review Board (IRB code: IR.MUMS.MEDICAL.REC. 1399.060.) of Mashhad University of Medical Sciences.

Table 1. Some of Clinical characteristics of patients admitted to infertility centers(IVF/ICSI).

Characteristic	Successful FHR (N=88)	Unsuccessful FHR (N=412)	Successful CPR (N=99)	Unsuccessful CPR (N=401)	Total (N=500)	*p*-value (CPR outcome)	*p*-value (FHR outcome)
Female Age	28.1 ± 5.2	29.5±5.4	27.9 ± 5.1	29.6±5.48	29.2±5.4	0.06[a]	0.02[a*]
FSH(mIU/ml)	7.4±2.8	6.6±3.2	7.3±2.7	6.6±3.2	6.8±0.4	0.11 [a]	0.04 [a*]

Table 2. Some of Clinical characteristics of patients admitted to infertility centers (IUI).

Characteristic	Successful FHR (N=68)	Unsuccessful FHR (N=161)	Successful CPR (N=78)	Unsuccessful CPR (N=151)	Total (N=229)	*p*-value (CPR outcome)	*p*-value (FHR outcome)
Female Age	29.7±5.4	29.7±5.4	31.3±6.4	31.02±5.8	31.3±5.9	0.09 [a]	0.6 [a]
FSH(mIU/ml)	8.15±5.3	8.15±5.3	8.7±10.7	8.3±4.5	8.52±6.2	0.1 [a]	0

Abbreviation. FSH: Follicle Stimulation Hormone, BMI: Body Mass Index, * Significant features (*p*-value <0.05), [a] Examined by student's t-test

2.2. Prediction Model

To establish the model, we used two well-known and Logistic Regression Model (LRM) as the non-parametric model and Random Forest Model(RFM) as a parametric model. After that, we evaluated the results by k-fold cross-validation (k=5) for the models. We validated the results by IUI dataset. Moreover, we obtained optimal cut-off for EMT by statistical test. Moreover, we described EMT categorized in terms of ongoing pregnancy and clinical pregnancy and the optimal cut-off for EMT.

3. Results

RFM predicted 89.9% and 79.4% from positive clinical pregnancy and positive ongoing pregnancy in patients under IVF/ICSI treatment and 82.6% and 73.5% in patients under IUI treatment, respectively. Also, RFM and LRM predicted ongoing pregnancy for IVF/ICSI treatment with, respectively Accuracy .69 and .65, Precision (PPV) .48 and .47 Recall .70 and .65, F-Score .57 and .55. Moreover, RFM and LRM predict ongoing pregnancy in IUI treatment, with Accuracy .82 and .80, PPV.77 and .67, Recall .82 and .81, F-Score .76 and .73. Accuracy values for models can be seen in Figure 1. A for IVF/ICSI and in Figure 1. C. for IUI. Area Under the Curve (AUC) values in the K-Fold process can be seen for IVF/ICSI in Figure 1. B. and in Figure 1.D. for IUI.

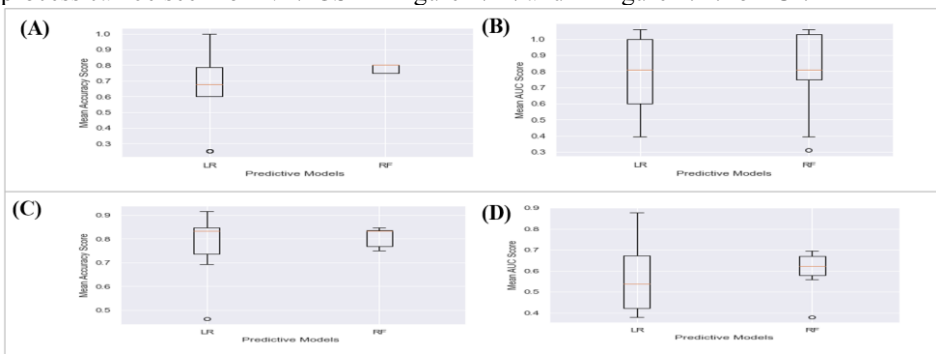


Figure 1. Comparison Accuracy and AUC respectively (A) and (B) for IVF/ICSI, (C) and (D) for IUI.

Also, we displayed relationship of EMT and pregnancy in Figure 2. Highest pregnancy rate (clinical and ongoing) is obtained 7.99-8.99 mm for IUI and 8.99-9.99 for IVF/ICSI treatment in Figure 2. A and Figure 2. B, respectively.

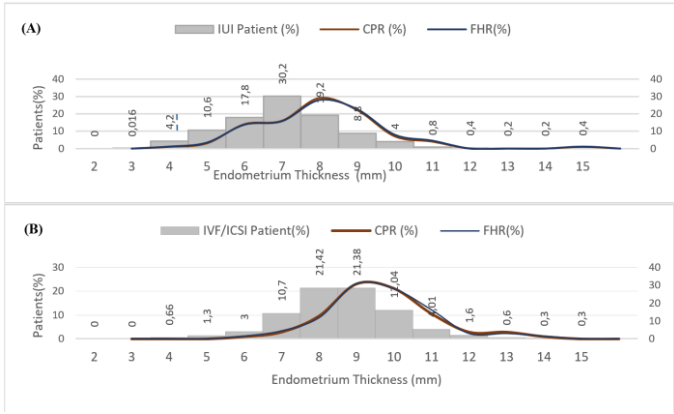


Figure 2. Relationship of Endometrium Thickness with CPR and FHR (A) for IUI; (B) for IVF/ICSI; Clinical pregnancy mean a pregnancy diagnosed by ultra-sonographic visualization of one or more gestational sacs. The clinical pregnancy rate (CPR) in IVF/ICSI is 34.6% (78/229), and for IUI is 19.8% (99/500).

Figure 3. A. and 3. B. indicated CPR and FHR with an error bar of 95% Confidence Intervals (CIs) by quantiles of EMT for both treatments. 22.86% and 20.15% of women undergoing IUI with EMT of 7 to 8.99 mm, obtained positive clinical and ongoing pregnancy, respectively (Figure 3. A.). Also, for IVF/ICSI treatment group, 36.56% and

33.58% of women with EMT of 8 to 9.99 mm obtained positive clinical and ongoing pregnancy, respectively (Figure 3. B.). EMT has associated with FHR based on Student T-test with (p<.001, 95% CI) for IUI and with (p=.024, 95% CI) for IVF/ICSI. Therefore, increasing EMT is statistically associated with increasing FHR before reaching the cut-off points. In addition, cut-off points for endometrial thickness have been calculated in both treatments based on Odd Ratio (OR) with a 95% confidence interval. Cut off points are obtained for IUI and IVF/ICSI treatment in 7.7 mm (OR=1.6, p-value=.05, CI: [1.08,2.7]) and 9.99 mm (OR=2.35, p-value=.03, CI: [1.05,5.9]) EMT, respectively. Statistically different is observed in EMT 9.99 mm with 4.99 – 8 mm in IVF/ICSI and EMT 7.7 mm with 3.99 – 7 mm in IUI.

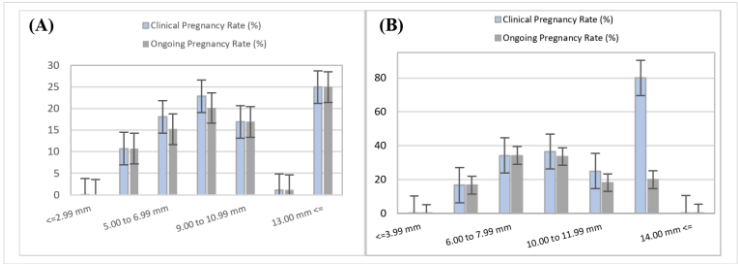


Figure 3. CPR and FHR by Percentile Endometrium Thickness in (A) for IUI and (B) for IVF/ICSI

Also, we analyzed output of the models by Shapley Additive exPlanations (SHAP) value plot [9]. The SHAP value helps interpret model prediction in terms of each feature. We draw the SHAP for RFM based on endometrial thickness for both of the treatment (See figure 4.). According to the results, endometrial thickness positively impacts RFM output for IVF/ICSI and IUI treatment

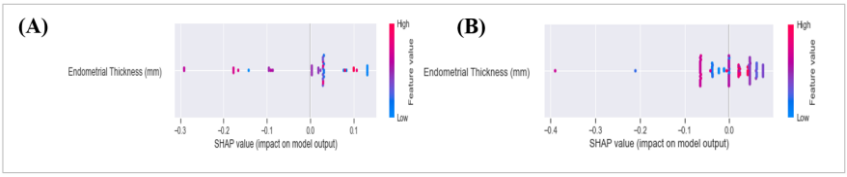


Figure 4. SHAP Value plots to show Impact of Endometrial Thickness on Model output; (A) for IVF/ICSI, (B) for IUI

4. Discussion

This is a retrospective study of 729 patients with unexplained infertility factors of IVF/ICSI and IUI treatment that led to the development of a machine learning-based model to predict the pregnancy rate. This study showed that RFM for prediction FHR and CPR have the best classification performance. In this study to examine pregnancy rate prediction, it can be seen that the use of machine learning tools can work better than the non-parametric regression model. Our study showed that clinical and ongoing pregnancy increase with increasing EMT until reaching the cut-off point. Although this increase was non-linear for both treatment methods. It showed a positive relationship between EMT and pregnancy rate. This correlation to pregnancy can be seen in similar studies [8; 9]. In contrast, some studies did not find a correlation between these two factor [12]. The fact that the study population is evaluated with different input and output

criteria can be influential. In both treatments after the cut-off, pregnancy is reduced, which may not be due to the negative effect of EMT on pregnancy. However, EMT on the day of HCG injection was a poor factor for both treatments; we obtained significant cut-off points of 7.7 and 9.99 mm in the IUI and IVF/ICSI treatments, respectively. While some similar studies obtained lower cut-off points for EMT [8; 9], some other studies obtained higher cut-off points by evaluating EMT factors per millimeters [5].

5. Conclusions

In summary, in contrast to LRM, machine learning-based models are more efficient and we found out machine learning is a valuable tool in predicting ongoing pregnancy and is trustable via multicenter data for two treatments Endometrial thickness was not statistically significantly different from CPR and FHR in both treatments. However, this difference was significant between patients with a positive FHR result in IUI and IVF/ICSI treatment.

References

[1] Practice Committee of the American Society for Reproductive Medicine. Effectiveness and treatment for unexplained infertility. Fertil Steril. 2006 Nov;86(5 Suppl 1):S111-4.
[2] Beall SA, DeCherney A. History and challenges surrounding ovarian stimulation in the treatment of infertility. Fertil Steril. 2012 Apr;97(4):795-801.
[3] Bromer JG, Aldad TS, Taylor HS. Defining the proliferative phase endometrial defect. Fertil Steril. 2009 Mar;91(3):698-704. doi: 10.1016/j.fertnstert.2007.12.066. Epub 2008 Mar 7.
[4] Esmailzadeh S, Faramarzi M. Endometrial thickness and pregnancy outcome after intrauterine insemination. Fertil Steril. 2007 Aug;88(2):432-7.
[5] Griesinger G, Trevisan S, Cometti B. Endometrial thickness on the day of embryo transfer is a poor predictor of IVF treatment outcome. Hum Reprod Open. 2018 Jan 29;2018(1):hox031.
[6] Habibzadeh V, Nematolahi Mahani SN, Kamyab H. The correlation of factors affecting the endometrial thickness with pregnancy outcome in the IUI cycles. Iran J Reprod Med. 2011 Winter;9(1):41-6. PMID: 25356081; PMCID: PMC4212145.
[7] Kim A, Young Lee J, Il Ji Y, Hyeog Lee H, Sil Lee E, Yeol Kim H, Oh YL. Do Endometrial Movements Affect The Achievement of Pregnancy during Intrauterine Insemination? Int J Fertil Steril. 2015 Jan-Mar;8(4):399-408. doi: 10.22074/ijfs.2015.4180. Epub 2015 Feb 7. PMID: 25780522; PMCID: PMC4355927.
[8] Liu KE, Hartman M, Hartman A, Luo ZC, Mahutte N. The impact of a thin endometrial lining on fresh and frozen-thaw IVF outcomes: an analysis of over 40 000 embryo transfers. Hum Reprod. 2018 Oct 1;33(10):1883-1888. doi: 10.1093/humrep/dey281. PMID: 30239738; PMCID: PMC6145412..
[9] Oron G, Hiersch L, Rona S, Prag-Rosenberg R, Sapir O, Tuttnauer-Hamburger M, Shufaro Y, Fisch B, Ben-Haroush A. Endometrial thickness of less than 7.5 mm is associated with obstetric complications in fresh IVF cycles: a retrospective cohort study. Reprod Biomed Online. 2018 Sep;37(3):341-348. doi: 10.1016/j.rbmo.2018.05.013. Epub 2018 Aug 23. PMID: 30146441.
[10] Quaas AM, Gavrizi SZ, Peck JD, Diamond MP, Legro RS, Robinson RD, Casson P, Christman GM, Zhang H, Hansen KR; Eunice Kennedy Shriver National Institute of Child Health and Human Development Reproductive Medicine Network. Endometrial thickness after ovarian stimulation with gonadotropin, clomiphene, or letrozole for unexplained infertility, and association with treatment outcomes. Fertil Steril. 2021 Jan;115(1):213-220. doi: 10.1016/j.fertnstert.2020.07.030. Epub 2020 Sep 21. PMID: 32972733; PMCID: PMC7796980.
[11] Smith S, Pfeifer SM, Collins JA. Diagnosis and management of female infertility. JAMA. 2003 Oct 1;290(13):1767-70. doi: 10.1001/jama.290.13.1767. PMID: 14519712..
[12] Wolff EF, Vahidi N, Alford C, Richter K, Widra E. Influences on endometrial development during intrauterine insemination: clinical experience of 2,929 patients with unexplained infertility. Fertil Steril. 2013 Jul;100(1):194-9.e1. doi: 10.1016/j.fertnstert.2013.03.023. Epub 2013 Apr 8. PMID: 23579008; PMCID: PMC3760031.

Challenges of Trustable AI and Added-Value on Health
B. Séroussi et al. (Eds.)

doi:10.3233/SHTI220451

Evaluating Unintended Consequences in Health Information Systems

Nurkhadija ROHANI[a] and Maryati Mohd YUSOF[b,1]
[a]*Pharmaceutical Policy & Strategic Planning Division, Ministry of Health, MALAYSIA*
[b]*Faculty of Info. Sc. and Technology, Universiti Kebangsaan Malaysia, MALAYSIA*

Abstract. The use of health information systems (HIS) and complex sociotechnical interactions can generate dangerous unintended consequences (UC). The evaluation of such interactions can provide an understanding of the root causes of UC. This paper reviews the interactions that lead to UC and its contributing factors.

Keywords: error, evaluation, framework, health information systems, patient safety, process, socio-technical, unintended consequences

1. Introduction

Health information systems (HIS) can improve healthcare quality [1,2]. However, HIS have various sociotechnical (social and technical) challenges limiting their potential and introducing new errors that may cause unintended consequences (UC). Healthcare providers can effectively understand their causes by recognising HIS-UC to mitigate them. This paper reviews HIS-UC factors and their complex interactions based on the Interactive Sociotechnical Analysis (ISTA) [3] and Human-Organization-Process-Technology-fit (HOPT-fit) frameworks [4].

2. Methods

We review the literature on UC issues arising from HIS implementation to determine the influencing factors. HIS problems are identified, classified, and summarised based on four ISTA interactions and mapped into HOPT factors in HIS.

3. Results

We identify 16 factors that influence UC in HIS from user perspectives and categorised them based on four selected ISTA interactions: (1) new HIS changes in the existing social system, (2) technical and physical infrastructures mediating HIS use, (3) the social system mediating HIS use and (4) the HIS-in-use changes in the social system. We also deduced the socio-technical dimensions for each UC (Table 1).

[1] Corresponding author; Maryati.Yusof@ukm.edu.my

Table 1. ISTA-UC-HOPT-fit framework interactions

▪ X = need fit, Y = evaluation dimension involved, – = not applicable		HOPT Dimension				
Interactions	**Factors Influencing UC**	**Org Structures**	**Workflow**	**System Use**	**Syst Quality**	**Info Quality**
ISTA 1: New HIS changes in the existing social system	1a – HIS disrupts current workflow and practice	-	Y	-	Y	-
	1b – HIS increases workload by changing current tasks and roles	Y	Y	Y	Y	-
	1c – HIS interferes with communication and information transfers	Y	Y	-	Y	-
	1d – HIS requires users to have sufficient knowledge and skills	-	-	Y	Y	-
ISTA 2: Infrastructures mediating HIS use	2a – HIS ability to integrate with existing infrastructure	Y	Y	Y	Y	-
	2b – Ability to provide technical support for HIS use	Y	-	-	Y	-
	2c – Availability of organisational infrastructure to support HIS use	Y	-	Y	-	-
ISTA 3: Social system mediating HIS use	3a – Incompatible HIS interfaces with practice add cognitive load	-	-	Y	Y	-
	3b – HIS functions are not interpreted according to the process	-	Y	Y	Y	Y
	3c – HIS functions are not well-adapted to communication and information transfers	-	Y	Y	Y	Y
	3d – Emergence of new types of errors	-	-	Y	Y	Y
ISTA 4: HIS-in-use changes in social system	4a – Changes in current practice	Y	Y	Y	Y	Y
	4b – Changes in HIS information content	Y	-	Y	Y	Y
	4c – Changes in communication and information transfers	Y	Y	Y	Y	Y
	4d – Overdependence on HIS	-	-	Y	Y	Y
	4e – Changes in users' awareness	-	-	Y	Y	Y
	Total	8	8	**13**	**15**	8

4. Discussion and conclusion

HIS use disrupts workflows (UC-1a), communication patterns and information transfers (UC-1c), confusing roles/responsibilities and unclear expectations (UC-1b) in the social system. HIS also transforms the physical and technical infrastructures (UC-2a) that require alignment with clinical workflows to support HIS use in the social system. However, multiple UCs can occur if the HIS functions (system quality) misaligned with the workflows (UC-3b) and communication and information transfers (UC-3c). HIS-UCs result from complex sociotechnical interactions that are difficult to understand. Therefore, evaluators can analyse them structurally to understand the interactions that can trigger HIS-UC for improvements and mitigation purposes.

Acknowledgement**.** Ministry of Higher Learning Malaysia (FRGS/1/2018/ICT04/UKM/02/5) sponsored the study.

References

[1] Dyb K, Warth LL.Implementing eHealth technologies: the need for changed work practices to reduce medication errors, in, IOS Press, 2019, pp. 83-86.
[2] Sittig DF, Wright A, Coiera E, et al. Current challenges in health information technology-related patient safety, Health Inform J 26 (2020), 181-189.
[3] Harrison MI, Koppel R, Bar-Lev S. Unintended consequences of information technologies in health care--an interactive sociotechnical analysis, JAMIA 14 (2007), 542-549.
[4] Yusof MM. A socio-technical and lean approach towards a framework for health information systems-induced error, in, IOS Press, 2019.

Challenges of Trustable AI and Added-Value on Health
B. Séroussi et al. (Eds.)

doi:10.3233/SHTI220452

Electronic Phenotyping to Identify Patients with Arrythmia Disease from a Hospital Information System

Masaharu NAKAYAMA[a,1] and Ryusuke INOUE[b]
[a] *Medical Informatics, Tohoku University Graduate School of Medicine, Miyagi, Japan*
[b] *Medical IT center, Tohoku University Hospital, Miyagi, Japan*

Abstract. Electronic phenotyping is an important method to identify a disease group by collecting clinical data from hospital information systems. This study aimed to extract accurate cases of supraventricular arrythmia, ventricular arrythmia, and bradycardia from clinical data of a hospital information system. The electronic phenotyping algorithm was improved using the machine learning method. Subsequently, it showed a higher area under the curve for prediction and higher specificity. However, the algorithm needs further improvement to classify each arrythmia disease accurately. In conclusion, phenotyping using clinical data from hospital information systems has some affinities and issues depending on the disease.

Keywords. Electronic phenotyping, machine learning, arrythmia disease.

1. Introduction

The use of medical big data is expected to provide real-world evidence research. Hospital information systems are very useful in this regard as they contain massive amounts of high-quality clinical data. Electric phenotyping has attracted attention in recent years as a tool to accurately identify patients diagnosed with a particular disease based on clinical data from hospital information systems [1,2]. This study attempted to extract true cases of several arrythmia diseases that can cause fatal conditions, such as strokes and sudden death because the algorithm to identify arrythmia diseases has not been established well.

2. Methods

The target diseases are arrythmias, including supraventricular arrythmia, ventricular arrythmia, and bradycardia. In this study, supraventricular arrythmia was defined as atrial extrasystole, atrial tachycardia, paroxysmal supraventricular tachycardia, atrial flutter, or atrial fibrillation. Ventricular arrythmia was defined as ventricular tachycardia, ventricular fibrillation, or ventricular extrasystoles. Bradycardia was defined as sinus failure syndrome and atrioventricular blocks. We used data comprising health insurance claims, diagnosis procedure combination data, and clinical information extracted from the Standardized Structured Medical Information eXchange version 2 standard storage

[1] Corresponding Author, Masaharu Nakayama, Japan; E-mail: nakayama@cardio.med.tohoku.ac.jp.

of Tohoku University Hospital (TUH) from April 1 to December 31, 2013 [3]. The initial rule was based on the International Classifications of Diseases 10th revision codes, examinations, therapeutic drugs, and procedures related to arrythmia diseases. Two physicians, including a cardiologist, checked the medical records to determine the true cases out of each 60 cases randomly selected; then, they calculated the positive predictive value (PPV). Using the evaluations, we trained the XGBoost model using five-fold cross-validation in R to improve algorithms, and then we evaluated the area under the curve (AUC) [4]. Finally, we retrospectively tested the algorithm on a cohort from an inpatient database of the TUH Department of Cardiovascular Medicine. This study was approved by the ethics review committee of Tohoku University (No. 2020-1-459).

3. Results

Regarding supraventricular arrythmia, although the initial rule had low PPV (27.8%), our model with XGBoost showed high correlation with radiofrequency catheter ablation, verapamil, and mexiletine (0.847 of AUC). The specificity was 96.0%. About ventricular arrythmia, the PPV was only 8.3% based on the initial rule. The applied method showed high correlation between implantable cardioverter-defibrillator (ICD), exchange of ICD, and amiodarone with 0.801 of AUC. The specificity was 97.4%, although the PPV increased to only 27.3%. For bradycardia, the PPV was initially 8.3%. The method showed high correlation with implantation and replacement of the pacemaker with 0.861 of AUC, although the PPV increased to only 23.3%. The specificity was 95.2%.

4. Discussion

Our study's results showed high specificity in the electronic phenotyping of arrythmia disease, resulting in usefulness of the related clinical data, such as disease names, examinations, medication, devices, and procedures, to exclude arrythmia cases among patients. However, there was little power to classify each arrythmia disease accurately because the use of related medication, examinations, and procedures overlapped among each arrythmia disease. This suggests that other important information, such as the Minnesota Code for electrocardiograms, would be critical to identify specific arrythmia cases. In conclusion, the findings suggest that phenotyping using clinical data from hospital information systems has some affinities and issues depending on the disease. To address these issues, it may be necessary to narrow down the target group more clearly based on the disease and collect disease-specific data.

References

[1] Shivade C, Raghavan P, Fosler-Lussier E, et al. A review of approaches to identifying patient phenotype cohorts using electronic health records. J Am Med Inform Assoc. 2014 Mar-Apr;21(2):221-230.
[2] Kashyap M, et al. Development and validation of phenotype classifiers across multiple sites in the observational health data sciences and informatics network. J Am Med Inform Assoc. 2020;27:877-83.
[3] Yamaguchi M, et al. Establishment of the MID-NET ® medical information database network as a reliable and valuable database for drug safety assessments in Japan. Pharma Drug Saf. 2019;28:1395-1404.
[4] Izukura R, et al. The development of an electronic phenotyping algorithm for identifying rhabdomyolysis patients in the MID-NET database. Stud Health Technol Inform. 2019 Aug 21;264:1498-1499.

Challenges of Trustable AI and Added-Value on Health
B. Séroussi et al. (Eds.)

doi:10.3233/SHTI220453

Utilizing Intensive Care Alarms for Machine Learning

Anne Rike FLINT[a], Sophie A. I. KLOPFENSTEIN[a,b], Patrick HEEREN[a], Felix BALZER[a] and Akira-Sebastian PONCETTE[a,1]
[a]*Institute of Medical Informatics, Charité – Universitätsmedizin Berlin, Germany*
[b]*Berlin Institute of Health, Charité – Universitätsmedizin Berlin, Germany*

Abstract. Alarms help to detect medical conditions in intensive care units and improve patient safety. However, up to 99% of alarms are non-actionable, i.e. alarm that did not trigger a medical intervention in a defined time frame. Reducing their amount through machine learning (ML) is hypothesized to be a promising approach to improve patient monitoring and alarm management. This retrospective study presents the technical and medical pre-processing steps to annotate alarms into actionable and non-actionable, creating a basis for ML applications.

Keywords. Alarm management, patient monitoring, machine learning

1. Introduction

Alarms are essential in a medical care setting as patient safety in intensive care units (ICU) relies on effective alarm systems and management [1]. Up to 99 % of alarms are non-actionable [2], meaning no medical intervention is required. So far, it is not possible to identify a medical intervention in a retrospective analysis of our hospital's alarm data nor do other publicly available databases contain such information. In this retrospective study, we aim to describe the requirements for a semi-automated annotation of alarm logs, enabling the deployment of machine learning (ML) algorithms at a later stage.

2. Methods

Through medical pre-processing, we identified clinical use cases, alarm types, patient health data related to interventions, and defined an annotation rule set based on conceptual mappings. During technical pre-processing, we extracted alarm logs from monitors and patient health data from the patient data management system, and explored database schema. We identified alarm start, pause and end times, and mechanisms to remap alarm log entries to patients' data. To execute the alarm annotation, we translated medical knowledge into executable scripts. IRB approval was obtained.[2]

[1] Corresponding Author, Akira-Sebastian Poncette, Institute of Medical Informatics, Charité – Universitätsmedizin Berlin, Charitéplatz 1, 10117 Berlin, E-mail: akira-sebastian.poncette@charite.de

[2]Ethics vote no. EA1/127/18, Ethikausschuss am Campus Charité Mitte, Chairperson: Prof. Dr.med. R. Morgenstern.

3. Results

For the timeframe March 2019 to October 2021, the alarm logs include 37,325 patients with 43,603 ICU stays. Clinical use cases and respective annotation rules are iterated and validated by focus groups with medical experts. “Low oxygen saturation” was chosen as the first use case, with an annotation rule set including numerical (e.g. oxygen saturation) and categorical (e.g. ventilation devices and modes) values. We applied the Medical Information Mart for Intensive Care (MIMIC) IV database schema [3,4] for patient health data with additional tables for alarm data. These include 40,942,722 alarm log entries with 15,551,660 alarm starts, 971,598 alarm pauses, and 15,555,489 alarm ends. Alarm starts are the reference point for annotations. The conceptual mappings are translated into a set of lookup tables. Python 3.7.6 and execute SQL statements automate the annotation and investigate the respective patient data points before and after an alarm.

4. Discussion

A database that combines patient health data and alarm data was built and a rule set for alarm annotation was defined. Alarm logs cannot be annotated in isolation, but rather in conjunction with other patient health data. The amount of medical and technical pre-processing presents a strong inhibitor to utilize ML for improved alarm management. Only a diverse team of medical and technical experts can give meaning to otherwise unlabeled, unstructured alarm data [5]. Manual annotations are cumbersome and error-prone [6]. Therefore, iterative implementation phases with quality insurance checks from clinicians can pave the way towards semi-automated alarm annotation.

5. Conclusion

A retrospective annotation of alarm data into actionable and non-actionable alarms could provide the basis for ML applications, pioneering a more intelligent, technology-aided or -driven alarm management.

References

[1] Sendelbach S, Funk M. Alarm fatigue: a patient safety concern. AACN Adv Crit Care. 2013 Oct-Dec;24(4):378-86; quiz 387-8. doi: 10.1097/NCI.0b013e3182a903f9. PMID: 24153215.
[2] Hravnak M, Pellathy T, Chen L, Dubrawski A, Wertz A, Clermont G, Pinsky MR. A call to alarms: Current state and future directions in the battle against alarm fatigue. J Electrocardiol. 2018 Nov-Dec;51(6S):S44-S48. doi: 10.1016/j.jelectrocard.2018.07.024. Epub 2018 Jul 29. PMID: 30077422
[3] Johnson AE, Pollard TJ, Shen L, Lehman LW, Feng M, Ghassemi M, Moody B, Szolovits P, Celi LA, Mark RG. MIMIC-III, a freely accessible critical care database. Sci Data. 2016 May 24;3:160035. doi: 10.1038/sdata.2016.35. PMID: 27219127.
[4] https://web.archive.org/web/20220126184518/https://mimic.mit.edu/docs/iv/, accessed 26.01.2022 19:45 Berlin, Germany.
[5] Wiens J, Shenoy ES. Machine Learning for Healthcare: On the Verge of a Major Shift in Healthcare Epidemiology. Clin Infect Dis. 2018 Jan 6;66(1):149-153. doi: 10.1093/cid/cix731. PMID: 29020316.
[6] Rad AB, Eftestøl T, Irusta U, Kvaløy JT, Wik L, Kramer-Johansen J, Katsaggelos AK, Engan K. An automatic system for the comprehensive retrospective analysis of cardiac rhythms in resuscitation episodes. Resuscitation. 2018 Jan;122:6-12. doi: 10.1016/j.resuscitation.2017.11.035. Epub 2017 Nov 6. PMID: 29122647.

Challenges of Trustable AI and Added-Value on Health
B. Séroussi et al. (Eds.)

doi:10.3233/SHTI220454

Modeling the Information Transparency of Health Service Privacy Policies

Marta ALIC[a] and Ljerka LUIC[b,1]
[a] *Zagreb University of Applied Sciences, Croatia*
[b] *University North, Croatia*

Abstract. In the present-day age of information abundance, in which the rights and freedoms in the digital environment are strengthened, information transparency is becoming an integral part of them. The rights of individuals to their own choice are more consequential in the field of privacy protection and the process of digital transformation in organizations is increasingly focused on the protection of collected and processed personal data. The basic (ex-ante) tool of transparency is the publication of privacy policies to inform the individuals with the procedures related to the collection, sharing, use and storage of their personal data, making them active shareholders in decision making process. The aim of this paper is to identify the factors influencing the information asymmetry of privacy policies in the field of health services and to provide a conceptual model for evaluating their information transparency.

Keywords. Information transparency, privacy policies, health service

1. Introduction

As the COVID-19 pandemy has disrupted the entire society safe, secure and standardized health information exchange has become imperative [1]. In such environments, it is difficult for users to identify risks to their personal data protection, especially since medical data, is considered to be sensitive data. The concept of information transparency is defined by drafting and publishing of privacy policy documents as mechanisms that contain all the information that users need to make informed decisions regarding their personal data. So, effective mechanisms for achieving transparency should aim to reduce information asymmetry between users and healthcare providers.

2. Methods

The ineffectiveness of transparency mechanisms can lead to infobesity or infoxication [2], a state in which more information is provided than is needed, leading to cognitive saturation or, conversely, to information starvation as a state in which less data is provided than needed. A state of overload or starvation arises when the depth of the information provided does not meet the requirements set in relation to the information, processes or policies provided.

[1] Corresponding Author, Ljerka Luic, University North, Trg dr. Zarka Dolinara 1, 48000 Koprivnica, Croatia; E-mail: ljerka.luic@unin.hr.

Depending on the content of the requirement and its category, it is necessary to set appropriate indicators that measure the depth of information for its fulfillment in relation to the ability of users to make decisions, while considering the services and organizational hierarchies of data processors, health service institutions. Audit of transparency mechanisms efficiency can be performed by the metrics proposed in [3] and in relation to taxonomic transparency requirements set by [4].

3. Results

To model efficient privacy policy mechanisms certain requirements on both dimensions of transparency should be considered: visibility, ie the degree of completeness of information and the possibility of finding them, and inferability, ie the degree to which information can be used to make the right decisions [5]. Dimension of visibility, focused on the content determinant of transparency, depends on indicators for informativeness, accessibility and currentness, while the degree of intervenability is characterized by qualitative characteristics of the mechanism of transparency itself, such as layering of the content, its meaningfulness and readability as indicators. By the ratio of the degrees of these two dimensions, it is possible to determine the level of information (a)symmetry in privacy policy documents.

4. Discussion

The proposed model for privacy policy information transparency evaluation sets the degree of information asymmetry as a reference value, which, in relation to the state of absolute symmetry, can be used to measure the values of elements on both dimensions of transparency.

5. Conclusions

By appointing information symmetry as a reference metric, the methodology for a standardized presentation of the information transparency factors within privacy policies can be developed and applied as a basis for risk assessment.

References

[1] Luic L, Striber-Devaja D. The significance of information standards for development of integrated health information system. Archive of Oncology. 2006;14(1-2):64-66.
[2] Hosseini M, Shahri A, Phalp K, Ali R. Engineering Transparency Requirements: a Modelling and Analysis Framework. Information Systems. 2017 Dec;74:3-38.
[3] Spagnuelo D, Bartolini C, Lenzini G. Metrics for Transparency. In: Livraga G, Torra V, Aldini A, Martinelli F, Suri N, editors. Data Privacy Management and Security Assurance. DPM QASA 2016 2016. Lecture Notes in Computer Science. Springer Cham; 2016. p. 3-18.
[4] Meis R, Wirtz R, Heisel M. A Taxonomy of Requirements for the Privacy Goal Transparency. In: Fischer-Hübner S, Lambrinoudakis C, López J, editors. Proceedings of the 12th International Conference, TrustBus; 2015 Sep 1-2; Valencia, Spain: Springer Cham; p. 195-209.
[5] Michener G, Bersch K. Identifying Transparency. Information Polity. 2013;18(3):233-242.

Challenges of Trustable AI and Added-Value on Health
B. Séroussi et al. (Eds.)

doi:10.3233/SHTI220455

Blockchain and IoT Technology in Healthcare: A Review

B SIVASANKARI[a,1] and P VARALAKSHMI[a]
[a]*Department of Computer Technology, Anna University, MIT Campus, Chennai, India*

Abstract. The tremendous shift in technology has led to many unconnected things getting interconnected via IoT. IoT is one of the major modes of collecting data from various networked resources and other connected devices. The broad range of IoT, with its huge heterogeneity in handling data, addresses many challenges in the realm of healthcare. Blockchain technology has elevated the use of distributed storage in a positive way. The recent emergence of this technology has paved way for potentially enormous utilization in various fields. Blockchain technology in the fields of IT, finance, industries, government, healthcare, media, and law enforcement has altered the service quality levels to an ethical ideal. Blockchain, in conjunction with IoT, facilitates decentralized collection and storage of data. Integrating blockchain with IoT has emerged as a cutting-edge tool for the decentralized sharing of medical records, monitoring of patients, ensuring the privacy of patient records, predicting the quantum of insurance, and managing supply chains.

Keywords. Blockchain, Internet of Things, Hyperledger Fabric, B-IoT (Blockchain-based IoT)

1. Introduction

IoT enables a network of physical entities by embedding sensors, operating software, and other technologies. IoT plays an important role in various fields due to its capability to provide sensor-based information and its ability to enable device-to-device communication. The advancement of sensing technologies has led to tremendous growth in data collection and data analysis.

Blockchain can be defined as a decentralized storage system that facilitates cryptographically secure storing and sharing of records immutably through a network comprising of distributed peer-to-peer members. The white paper laid the foundation for bringing solutions to globally connected technology, comprising of decentralized monetary mechanisms [1]. In this paper, they describe how Bitcoin evolved and took technology deployment in a decentralized network to the next level.

[1] B Sivasankari, Corresponding author, Department of Computer Technology, MIT Campus, Anna University, Chennai – 600044, India; E-mail: sivashankaripitam@gmail.com

2. IoT and Blockchain in Healthcare

IoT layer is made of sensors, motors, controllers, gateways, and other similar computational resources. This layer also facilitates the storage and management of these resources. It uses wireless standards such as ZigBee, Wi-Fi, or LoRa [2]. The edge layer serves the purpose of coordination of different technical assets and it is close to the end-users.

The application layer consists of elements related to cloud management, data analysis, authorization, and knowledge base. It has a collection of services and business applications. Patient's Electronic Health records, present and past complaints, day-to-day routines, medication details, and genomic data can be measured using IoT sensors [2]. The Blockchain rectifies the drawbacks of the centralized storage system and cooperates with edge computations to securely store information on the network. A Hyperledger platform is a permissioned distributed ledger for deploying applications with access control specifications [3].

3. Applications of Blockchain-based IoT (B-IoT) in Health Care

A consortium blockchain-based medical information sharing system uses k-anonymity, and keyword searchable encryption to ensure data privacy and security. It scrutinizes every single payment and lowers the load on insurance auditors. The layered architecture of B-IoT comprises a physical layer, blockchain layer, and application layer. The physical layer consists of sensor nodes as end devices of the IoT network. Data aggregated from the IoT network are stored in the blockchain layer securely in a decentralized manner. Applications are executed in the application layer by retrieving the data from the blockchain layer.

4. Conclusion

This survey examines the various aspects of the literature pertaining to bundling IoT with blockchain in the healthcare domain. The decentralized storage system used in blockchain with IoT guarantees privacy and security in sharing critical data. Important Sensors in IoT networks and their usage in the health sector have been described. The future of Blockchain rests on integrating machine learning and the Internet of Things with Blockchain technology to develop an autonomous network with distributed storage.

References

[1] Nakamoto S. Bitcoin: A Peer-to-Peer Electronic Cash System, 2008, Available: http://bitcoin.org/bitcoin.pdf.

[2] ElRahman SA, Alluhaidan AS. Blockchain technology and IoT-edge framework for sharing healthcare services, Springer, Soft Computing, 2021, Nov; 25:13753–13777.

[3] Xu X, Wang X, Li Z, et al. Mitigating Conflicting Transactions in Hyperledger Fabric-Permissioned Blockchain for Delay-Sensitive IoT Applications, IEEE Trans. IEEE Internet of Things Journal, 2021, Jul; vol. 8, no. 13, pp. 10596 – 10607.

Challenges of Trustable AI and Added-Value on Health
B. Séroussi et al. (Eds.)

doi:10.3233/SHTI220456

Intra-Query Parallelism for a Scalable and Responsive Web-Based Digital Pathology Viewer

Rui JESUS[a,1], Luís Bastião SILVA[b,c] and Carlos COSTA[b]
[a] *University of A Coruña, Spain*
[b] *University of Aveiro, Portugal*
[c] *BMD Software, Portugal*

Abstract. Computational systems are successfully used in distinct areas of health. However, the use of digital imaging in pathology is taking its first steps when compared with the radiology sector. For instance, inter-institutional web platforms interoperable with various scanners through standard communications are very rare. This is due to the fact that the technology is only now being mature enough to meet the major challenges of this sector in terms of data storage and data access for efficient visualization of images with several gigabytes. The remote access to those images proves to be a challenge in an open and heterogeneous environment. This paper proposes a scalable and efficient architecture for storing and dynamic data retrieval on distributed large-scale systems. The adopted methodology relies on intra-query parallelism to retrieve a large number of image segments in a scalable distributed environment.

Keywords. WSI, Digital Pathology, PACS, Intra-Query Parallelism, Distributed systems

1. Introduction

The presence of digital technologies in the healthcare industry has become more prevalent due to the advantages that it brings over traditional ones. Improvements in workflow speed and ergonomics [1], quality of the diagnosis, access to diagnostic support tools, collaborative sessions and management of the medical data were made possible by storing the medical exams, and scans in a digital format [2].

Despite the advantages digital pathology brings, these systems are still relatively new, and their adoption has been slow. The implementation of a digital pathology system in a large-scale clinical center poses several challenges. It needs to provide significant advantages in both operation costs and logistics and most importantly, it needs to improve diagnose efficiency and quality to justify the migration efforts [3]. This is a considerable challenge for web-based systems in digital pathology due to the giga-pixel resolution of the digital slites.

This paper proposes a distributed architecture that takes advantage of the data organization in a Whole Slide Image (WSI) to increase system performance and reliability in large-scale scenarios.

[1] Corresponding Author, Campus Universitário de Santiago, Aveiro, Portugal; E-mail: r.jesus@udc.es

2. Architecture Overview

The goal of this architecture (Figure 1) is to reduce response times to user's queries when the system is under heavy load but also to singular query requests. The user navigates the WSI images through the WSI Viewer component. These images can be seen as layers of a pyramid split in tiles. A slide scanned at 40x magnification can have upwards of 30000 tiles hence the challenge that they pose. When the user navigates the image, several tiles will be requested at once to update the application viewport. Each frame is individually requested through a HTTP request. This allows the architecture to split the workload through the node network employing an intra-query strategy. The Server Application splits the requests equally amongst all nodes using load balancing strategies. All nodes can answer image requests but only the master node is allowed to receive data. This node propagates the data it receives to the remaining nodes in the network.

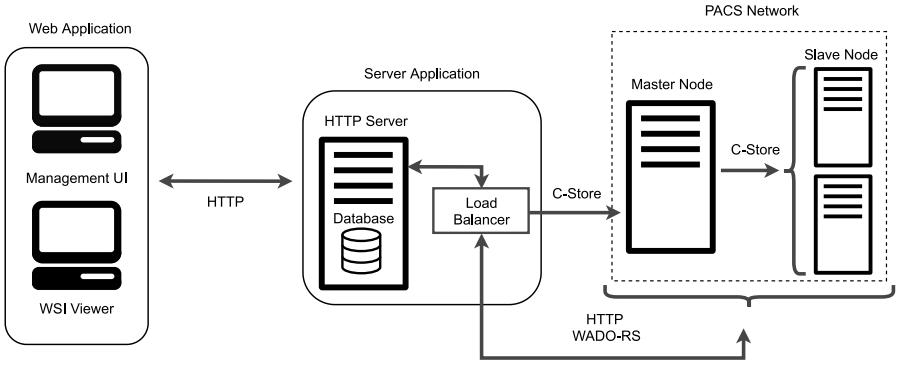


Figure 1. Overview of the proposed system architecture

3. Conclusion

The architecture presented in this work tries to address the concerns of a digital pathology system being used on a massive scale over the Web. It was proposed an architecture to facilitate access to the tiles of a WSI in a parallelized format, using an intra-query strategy. The results show that by splitting the work through several archive nodes response times can be improved with the proposed architecture.

This work has received support from iPATH project (POCI-01-0247-FEDER-047069).

References

[1] Vodovnik A, Diagnostic time in digital pathology: A comparative study on 400 cases, J. Pathol. Inform., 2016 Jan;7(1)

[2] Lebre R, Jesus R, Nunes P, Costa C. Collaborative framework for a whole-slide image viewer, in Proceedings - IEEE Symposium on Computer-Based Medical Systems. 2019 jun.

[3] SW J, M P, F M, Digital Pathology: Advantages, Limitations and Emerging Perspectives, J. Clin. Med., 2020 nov;9(11):3697.

Challenges of Trustable AI and Added-Value on Health
B. Séroussi et al. (Eds.)

doi:10.3233/SHTI220457

SOIN – *MI Data Lab*: Personalized Ophthalmology Through Collaborative Data Collection and Dynamic Patient Consent

Ciara BERGIN [a], Irmela MANTEL [a], Reinier O. SCHLINGEMANN [a], Mattia TOMASONI [a], Thomas J. WOLFENSBERGER [a,1]
[a] *Jules-Gonin Eye Hospital, Lausanne, Switzerland*

Abstract. The *Swiss Ophthalmic Image Network* (SOIN) is part of the *Swiss Personalized Health Network* (SPHN). SOIN contains a collaborative, clinical research environment, *MI Data Lab*, which allows privacy-preserving, data-driven, research. Personalized care of chronic ocular disease, based on Machine Learning (ML) and medical imaging, can dramatically improve quality of life and reduce the burden on health and social care systems. *MI Data Lab* allows research partners to consolidate their data in a space where doctors and data scientists cooperate to design novel ML algorithms, on curated datasets. To date, we have created several algorithms to detect ocular biomarkers automatically, and applied such tools to 100k+ retinal images. *MI Data Lab* enables the development of predictive models, the extraction novel traits to be explored in terms of -omic associations, treatment outcome, and priors for disease progression.

Keywords. Personalized medicine, research platform, dynamic patient consent, ophthalmology, SOIN, MI Data Lab, scientific IT services, sensitive data

1. Introduction

Machine Learning (ML) algorithms hold the potential to revolutionize personalized care of chronic ocular diseases. Research in ophthalmology is often hindered by proprietary formats, lack of dynamic patient consent, and difficulty of extracting complementary variables from patient records. These factors significantly hinder the creation of adequate training datasets for ML. The aim of the SOIN project is to address these issues by providing a collaborative space where data from consenting patients is seamlessly aggregated into study-specific datasets.

2. Methods

The *Swiss Ophthalmic Image Network* (SOIN) is part of the *Swiss Personalized Health Network* (SPHN). *MI Data Lab* (Medical Imaging Data Lab) is a collaborative, clinical research environment created during the SOIN project. *MI Data Lab* was built within BioMedIT, a secure infrastructure provided by the SPHN [1]. BioMedIT offers technical

[1] Corresponding Author: Thomas J. Wolfensberger, University of Lausanne, Jules-Gonin Eye Hospital, Fondation Asile des aveugles, Av. de France 15, 1004 Lausanne, E-mail: thomas.wolfensberger@fa2.ch

support for the onboarding of new users and the customization of user-specific spaces. Access to *MI Data Lab* is cloud-based, with double authentication, resolving accessibility and interoperability issues.

MI Data Lab, is a closed space, with no internet connectivity, where data import/export are technically and judrically restricted (in line with governance principles for data reuse). A unidirectional migration of medical data, images, and consent status has been established between the Jules-Gonin hospital's IT infrastructure and *MI Data Lab*. Data are transformed into open formats during the transfer process. A data dictionary is supplied for additional data fields, such as derived patient phenotypes. The data required are specified during project submission, and they are unpacked in a project-specific workspace: this way, basic data engineering tasks and common software needs are taken care of. Data access and user account management are performed centrally, per project. The iterative process of enriching available tools, software packages, and environments, continues to be the focus of the SOIN project.

3. Results

The proposed paradigm eliminates the need for a data transfer agreement, shortening the time from project conception to development, from years to weeks. To date, *MI Data Lab* has been used to extract ocular biomarkers from 100k+ retinal images, demonstrating its effectiveness.

4. Discussion

Data remain consultable by the clinician thanks to an embedded viewing software. The next objective of the SOIN project is to allow manual annotation tasks to be queued, and subsequently made available for review.

5. Conclusion

MI Data Lab facilitates rapid access to sensitive patient data to approved researchers, in a format and environment adapted to ML development needs. MI Data Lab aims to accelerate the development of predictive models and decision support systems, improving the quality/cost ratio of ophthalmic care and making important gains in personalized medicine achievable in ophthalmology.

References

[1] Coman Schmid D, Crameri K, Oesterle S, Rinn B, Sengstag T, Stockinger H, et al. SPHN - The BioMedIT Network: A Secure IT Platform for Research with Sensitive Human Data. Stud Health Technol Inform. 2020;270: 1170–1174.

Challenges of Trustable AI and Added-Value on Health
B. Séroussi et al. (Eds.)

doi:10.3233/SHTI220458

Microservices-Based Architecture to Support the Adaptive RECORDS-Trial

Ebrahim NAGEBA[a,1] Martin HILKA[a] Rafael GOZLAN[a] Julien DUBIEL[a], Cendrine BAUDOIN[a], Christel DANIEL[a]
[a] *AP-HP, Information Systems Direction, Innovation and Data Department*

Abstract. Information systems used by platform trials should handle changes that are not predefined. Unfortunately, the technical architecture of most existing clinical data management systems (CDMS) do not support changes to be incorporated into an ongoing trial. Adaptive clinical trials need advanced architectural solutions setup to enable biomarker stratification and enrichment strategy necessary for the adaptive clinical trial operation. This short paper presents the microservices-based architecture solution that is used to run and support the adaptive RECORDS-Trial.

Keywords. Platform Trials, Adaptive Clinical-Trial, Biomarker stratification, Microservices-based Architecture, Machin-Learning Algorithms, Interoperability, Containerization, Clinical Data, CDMS, eCRF.

1. Introduction

Adaptive clinical trials incorporate multiple comparisons in the context of a disease or treatment. Changes are performed during the conduct of the trial in response to patient information accumulating and adaptive elements using pre-defined interim analyses [1]. In some platform trials, treatment assignments are based on the patients' biomarker profiles and precision health methods are incorporated into the interim and final analyses. Changes may need to integrate new biomarkers modifying the randomization process. Such changes should not impact the clinical study infrastructure or the clinical trial operation. With the emergence of AI-based biomarkers, biomarker stratification at the point of care is technically especially challenging. In the next section, we briefly describe the technical architecture solution used to integrate predictive algorithms and randomization application of the adaptive RECORDS-Trial.

2. Microservices-based architecture to support REDCORDS-Trial.

The Adaptive Clinical-Trial RECORDS (Rapid rEcognition of CS sensitive or resistant Sepsis) [2], needs robust, flexible, and interoperable technical architecture to integrate machine-learning algorithms developed by RECORDS project partners (i.e Versailles University, Berkeley University) to run AI-based biomarker stratification [2]. The trained algorithms are used to automatically predict features starting from dataset entered in inclusion form within the eCRF application. Predicted features are essential to

[1] Ebrahim Nageba, Campus Picpus, Innovation and Data department, ebrahim.nageba@aphp.fr

determine patient the treatment options for specific patients, during the conduct of RECORDS-Trail.

To enable the adaptive RECORDS-Trial to use trained predictive algorithms, the Data and Innovation department has set up a microservices-based, extensible, and interoperable technical architecture. Each machine-learning algorithm, developed by RECORDS project partners, is wrapped by web framework to be used as a web service (WS). For instance, Flask framework has been used to wrap algorithms developed using Python Programming Language, and Plumber framework has been used to wrap algorithms developed using R Language. The web-based algorithms then have been containerized using Docker technology. On the other hand, the eCRF application used by adaptive RECORDS-Trial has integrated the WS related to the predictive algorithms in order to be called at each patient inclusion in RECORDS Interventional Study.

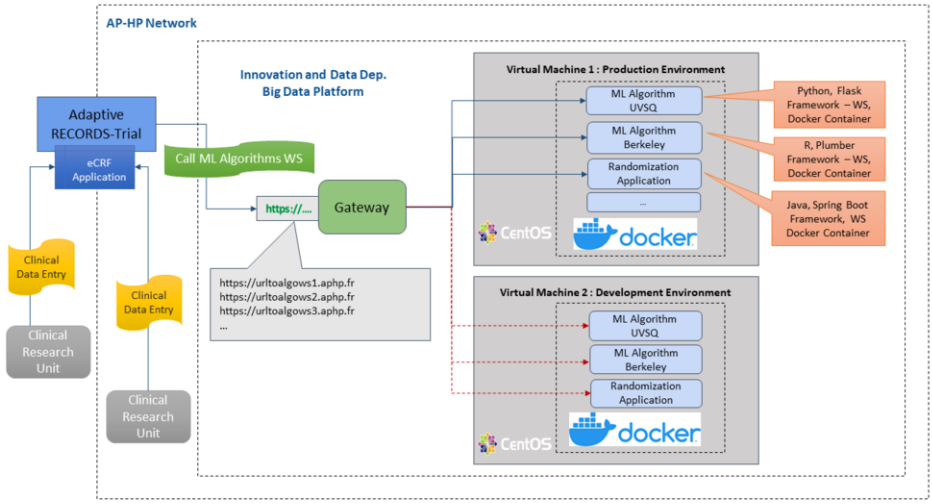


Figure 1. Global overview of the Microservices-based architecture to support the adaptive RECORDS-Trial

CDMS that integrates API (Application Programming Interface) can effectively interact with microservices and exchange data needed by adaptive clinical trial. The schema above shows an overview of the microservices-based architecture currently used to support the adaptive RECORDS-Trial. Open Source technologies like CentOS Operating System, Docker Engine, and Docker-Compose, are used to create the microservices architecture. To secure the microservices, all data streams are encrypted using a security certificate. In addition, HTTPS protocol is used to query the WS, as well as, all data flows pass by a gateway that uses additional security components such as Nginx Revers Proxy.

References

[1] Berry SM, Connor JT, Lewis RJ. The Platform Trial An Efficient Strategy for Evaluating Multiple Treatments, JAMA, volume 313, number 16, 2018, p.1619-1620.

[2] Pirracchio R, Hubbard AL, Sprung C, Chevret S, Annane D. Assessment of Machine Learning to Estimate the Individual Treatment Effect of Corticosteroids in Septic Shock, 2020, DOI: 10.1001/jamanetworkopen.2020.29050

Section III

Knowledge and Information Representation and Modeling

This page intentionally left blank

Challenges of Trustable AI and Added-Value on Health
B. Séroussi et al. (Eds.)

doi:10.3233/SHTI220460

Building an i2b2-Based Population Repository for COVID-19 Research

Miguel PEDRERA-JIMENEZ[a,b,1], Noelia GARCIA-BARRIO[a], Gema HERNANDEZ-IBARBURU[c], Blanca BASELGA[a], Alvar BLANCO[a], Fernando CALVO-BOYERO[a], Alba GUTIERREZ-SACRISTAN[d], Víctor QUIROS[a], Juan Luis CRUZ-BERMUDEZ[a], José Luis BERNAL[a], Laura MELONI[e], David PEREZ-REY[c], Matvey PALCHUK[d,e], Isaac KOHANE[d] and Pablo SERRANO[a]

[a]*Data Science Unit, Research Institute Hospital 12 de Octubre, Madrid, Spain*
[b]*ETSI Telecomunicación, Universidad Politécnica de Madrid, Madrid, Spain*
[c]*Biomedical Informatics Group, Universidad Politécnica de Madrid, Madrid, Spain*
[d]*Department of Biomedical Informatics, Harvard Medical School, Boston, MA, USA*
[e]*TriNetX, LLC, Cambridge, MA, USA*

Abstract. Reuse of Electronic Health Records (EHRs) for specific diseases such as COVID-19 requires data to be recorded and persisted according to international standards. Since the beginning of the COVID-19 pandemic, Hospital Universitario 12 de Octubre (H12O) evolved its EHRs: it identified, modeled and standardized the concepts related to this new disease in an agile, flexible and staged way. Thus, data from more than 200,000 COVID-19 cases were extracted, transformed, and loaded into an i2b2 repository. This effort allowed H12O to share data with worldwide networks such as the TriNetX platform and the 4CE Consortium.

Keywords. Electronic Health Records, Real World Data, Data Reusability, Semantics, Standardized repositories, i2b2, TriNetX, 4CE Consortium, COVID-19.

1. Introduction

COVID-19 pandemic has been the major health challenge in recent decades, being declared a pandemic state on March 11, 2020 by World Health Organization (WHO) [1]. In this critical situation, Electronic Health Records (EHRs) have proven crucial for patient management and healthcare, as well as for clinical research on this new and unknown disease [2].

In this sense, the reuse of EHRs for COVID-19 research required health data to be available in an agile way, flexible to specific requirements, standardized according to international recommendations and exploitable through advanced analysis tools [3]. Some of the health data resources that facilitate this task are the standardized clinical repositories such as i2b2 and OMOP CDM [4, 5].

Thus, this work aims to describe how Hospital Universitario 12 de Octubre in Madrid, Spain (H12O) has developed and used its i2b2 repository [6] to respond to the data needs that arose during the COVID-19 pandemic.

[1] Corresponding Author, Miguel Pedrera Jiménez, Data Science Group, Hospital Universitario 12 de Octubre, Av. de Andalucía S/N, 28041, Madrid, España; E-mail: miguel.pedrera@salud.madrid.org.

2. Methods

This work was carried out at H12O, implementing this organization's methodology to effectively reuse EHRs for research [3, 7]. The sources and type of data required, standardization resources, mapping efforts, and data volumes obtained are provided in this paper to help other healthcare institutions willing to share COVID-19 data.

2.1. EHRs from Hospital Universitario 12 de Octubre

The source for data extraction was the healthcare information systems of H12O, which have been formally modeled through health information standards such as ISO 13606, and incorporate terminologies such as SNOMED CT and LOINC. This standardization of the EHRs has allowed their reuse, without additional manual efforts and full meaning, in data collection processes for research and other secondary uses [3].

To make EHRs reuse efficient, we designed a set of multipurpose information models, common to all use cases. These were formalized through clinical archetypes, and implemented in the healthcare information systems of the hospital. Hence, these standardized information resources have been used to implement research tools at H12O, including a clinical repository based on the i2b2 data model [6]. Table 1 describes this set of information models for EHRs reuse.

Table 1. Information models for EHRs reuse of Hospital Universitario 12 de Octubre.

Archetype	Description	Terminology binding
Patient	Demographics data, e.g., birthdate, sex and vital status.	SNOMED CT
Encounter	Data related to inpatient, emergency and outpatient visits.	SNOMED CT
Location	Patient locations during hospitalization, e.g., ICU admission.	SNOMED CT
Observation	Clinical, laboratory and patient-reported observations.	SNOMED CT, LOINC
Diagnosis	Health issues and clinical diagnoses.	SNOMED CT
Medication	Pharmacological treatment prescribed.	SNOMED CT
Procedure	Procedures performed, e.g., surgeries and nursing interventions.	SNOMED CT

2.2. Informatics for Integrating Biology & the Bedside (i2b2)

The i2b2 tool is a scalable informatics framework that organizes and transforms patient-oriented health data in a way that is optimized for research. It was developed by Harvard Medical School with funding from the National Institutes of Health (NIH).

The database design for the clinical repository is based on a star schema composed of a central table for the observed clinical facts, which is related to five additional tables that provide contextual information about the observation, i.e., patient, visit, provider, and observable entity. Figure 1 shows the data model proposed by i2b2.

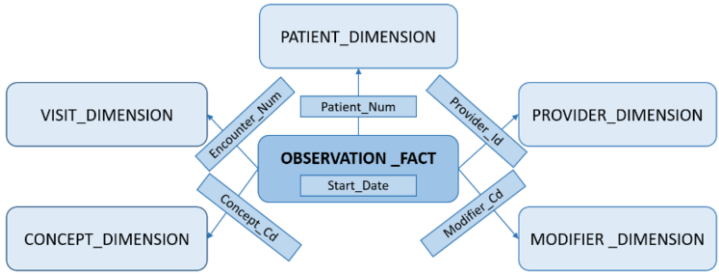


Figure 1. Data model for clinical repository proposed by i2b2.

2.3. COVID-19-related concepts

COVID-19 meant facing a new and unknown disease, so healthcare information systems had to be evolved, in an agile, flexible and staged way, to include these new concepts necessary for diagnosis, treatment and prevention of this condition [2].

This set of concepts was identified through a group of clinical domain experts at H12O. For this purpose, we adopted the WHO recommendations for patient stratification during the pandemic [8], and the Spanish Ministry of Health terminology standards [9]. Table 2 shows the set of COVID-19-related concepts that were modeled and standardized, indicating the terminology, and the date when these concepts were implemented into healthcare information systems.

Table 2. Standardized set of COVID-19-related concepts of H12O ordered by implementation date.

EHRs concept	Description	Since
SNOMED-CT: 63681000122103	Diagnoses disease caused by 2019 novel coronavirus with specific diagnostic tests	2020-03-08
SNOMED-CT: 840544004	Suspected disease caused by 2019 novel coronavirus	2020-03-08
SNOMED-CT: 63341000122104	Type of respiratory support	2020-03-08
SNOMED-CT: 704296008	At risk of impaired respiratory system function	2020-03-08
LOINC: 94315-9	SARS-related coronavirus E gene [Presence] in Specimen by NAA with probe detection	2020-03-17
SNOMED-CT: 62811000122102	Diagnoses disease caused by 2019 novel coronavirus without specific diagnostic tests	2020-03-17
SNOMED-CT: 160734000	Lives in a nursing home	2020-03-20
SNOMED-CT: 688232241000119100	Discarded disease caused by 2019 novel coronavirus	2020-04-15
SNOMED-CT: 736534008	EuroQol five dimension five level index value (observable entity)|	2020-06-29
LOINC: 94558-4	SARS-CoV-2 (COVID-19) Ag [Presence] in Respiratory specimen by Rapid immunoassay	2020-09-29
LOINC: 96751-3	SARS-CoV-2 (COVID-19) S gene mutation detected [Identifier] in Specimen by Molecular genetics method	2020-12-23
LOINC: 96895-8	SARS-CoV-2 (COVID-19) lineage [Identifier] in Specimen by Molecular genetics method	2020-12-23
SNOMED-CT: 1156257007	Administration of vaccine product against Severe acute respiratory syndrome coronavirus 2	2021-01-10

3. Results

Our main results are to put in place a whole extraction, transformation, and loading process (ETL) of the patient-level COVID-19 related data into an i2b2 repository and the successful usage for research across multiple projects.

3.1. Extraction, transformation and loading of COVID-19 data

First, the data was extracted, transformed (operation T.1.2, change of coding system [7]), and loaded into the i2b2 repository. Table 3 shows, for each concept related to COVID-19, the terminology mapping implemented, the volume of records and the number of patients loaded on January 17, 2022.

Table 3. Volume of records and patients related to COVID-19 loaded into the i2b2 repository.

EHRs concept	Mapped concepts	Records (N)	Patients (N)
SNOMED-CT: 63681000122103	ICD10CM:U07.1	56,051	53,300
SNOMED-CT: 840544004	ICD10CM:Z20.822	9697	8415
SNOMED-CT: 63341000122104	ICD10PCS:5A09, ICD10PCS:5A19	8274	7835
SNOMED-CT: 704296008	*No mapping required*	7427	7026
LOINC: 94315-9	LOINC: 94315-9	355,432	211,656
SNOMED-CT: 62811000122102	ICD10CM:U07.2	1402	1307
SNOMED-CT: 160734000	ICD10CM:Y92.12	2277	2240
SNOMED-CT: 688232241000119100	ICD10CM: Z03. 818	24,386	22,084
SNOMED-CT: 736534008	LOINC: 97332-1	3250	1184
LOINC: 94558-4	LOINC: 94558-4	46,850	35,441
LOINC: 96751-3	LOINC: 96751-3	1124	1094
LOINC: 96895-8	LOINC: 96895-8	13,906	13,152
SNOMED-CT:1156257007	RXNORM:2468231	279,450	160,738

3.2. Use of COVID-19 data for research

COVD-19 data stored into the H12O i2b2 repository have been used in several projects, including two international ones, the TriNetX platform and the 4CE Consortium.

3.2.1. TriNetX platform

The TriNetX platform is a global health research network to share real-world data, making clinical and observational research more accessible and efficient [10]. This platform combines real-time access to longitudinal clinical data with state-of-the-art analytics to optimize protocol design and feasibility, site selection, patient recruitment, and enable discoveries through the generation of real-world evidence. It contains data from more than 120 sites in 19 countries.

Data on more than 200,000 COVID-19 cases (suspected, diagnosed, confirmed and discarded) were loaded in real-time as of March 8, 2020. This allowed us to respond to questions of general interest from the beginning of the pandemic. One of the most important was to determine the impact on the patient outcome of coming to the hospital from a nursing home [11]. Similarly, this tool has allowed the development of complex clinical studies of high impact, combining COVID-19 data with data from other conditions. Among other initiatives, a relevant study was developed by H12O Hematology Department, which confirmed that the COVID-19 pandemic has a more severe impact on patients with Multiple Myeloma (MM) than non-MM patients [12].

3.2.2. 4CE Consortium

The 4CE Consortium federates more than 300 hospitals from seven countries, and so far, the H12O is the only Spanish hospital [13]. This was possible thanks to the joint efforts from the i2b2 community and the health data experts from H12O. Thus, having the data in the i2b2 repository and having as reference the extraction and transformation scripts developed by the 4CE Consortium, we could obtain the required data in a timely manner.

Hence, we could extract and aggregate patient-level data from 7,028 COVID-19 cases. This work, and the international effort of this consortium, allowed doing relevant research on patients affected by COVID-19. In one of these, data from 671 children hospitalized were included in a pediatric study, being 78 of them from H12O [14].

4. Conclusions

The rapid identification, modeling and standardization of COVID-19 concepts into EHRs allowed obtaining valuable data for research on this new disease. These data were loaded in an agile, flexible and staged manner into the i2b2 repository of H12O for having data available according to standards and exploitable by advanced analysis tools.

Thus, the TriNetX platform was used to answer research questions since the beginning of the COVID-19 pandemic. Similarly, H12O joined the 4CE Consortium, combining data with other hospitals, and participating in relevant international studies.

Acknowledgment

This work has been supported by Research Projects PI18/00981, PI18/01047 and PI18CIII/00019; funded by Instituto de Salud Carlos III, co-funded by ERDF/ESF.

We want to thank TriNetX and 4CE Consortium for their support in this project.

References

[1] Zhu N, Zhang D, Wang W, et al. A novel coronavirus from patients with pneumonia in China, 2019. N Engl J Med 2020;382:727–33. doi:10.1056/NEJMoa2001017.

[2] Pedrera M, Garcia N, Blanco A, et al. Use of EHRs in a Tertiary Hospital During COVID-19 Pandemic: A Multi-Purpose Approach Based on Standards. Stud Health Technol Inform. 2021;281:28-32. doi:10.3233/SHTI210114.

[3] Pedrera-Jiménez M, García-Barrio N, Cruz-Rojo J, et al. Obtaining EHR-derived datasets for COVID-19 research within a short time: a flexible methodology based on Detailed Clinical Models. J Biomed Inform. 2021;115:103697. doi:10.1016/j.jbi.2021.103697.

[4] Murphy SN, Weber G, Mendis M, et al. Serving the enterprise and beyond with informatics for integrating biology and the bedside (i2b2). *J Am Med Inform Assoc*. 2010;17(2):124-130. doi:10.1136/jamia.2009.000893.

[5] Hripcsak G, Duke JD, Shah NH, et al. Observational Health Data Sciences and Informatics (OHDSI): Opportunities for Observational Researchers. Stud Health Technol Inform. 2015;216:574-578.

[6] González L, Pérez-Rey D, Alonso E, et al. Building an I2B2-Based Population Repository for Clinical Research. *Stud Health Technol Inform*. 2020;270:78-82. doi:10.3233/SHTI200126.

[7] Pedrera M, Garcia N, Rubio P, Cruz JL, Bernal JL, Serrano P. Making EHRs Reusable: A Common Framework of Data Operations. Stud Health Technol Inform. 2021;287:129-133. doi:10.3233/SHTI210831.

[8] WHO ICD-10 codes for COVID-19. https://www.who.int/classifications/icd/COVID-19-coding-icd10.pdf. Accessed January 17. 2021.

[9] Health Ministry of Spain: Minimum data set for clinical reports. https://www.boe.es/eli/es/rd/2010/09/03/1093. Accessed January 17. 2021.

[10] TriNetX Platform. https://trinetx.com/. Accessed January 17. 2021.

[11] COVID-19 and nursing homes at H12O. https://www.comunidad.madrid/noticias/2021/12/16/hospital-12-octubre-premio-barea-su-trabajo-centros-socio-sanitarios. Accessed January 17. 2021.

[12] Martinez-Lopez J, Hernandez-Ibarburu G, Alonso R, et al. Impact of COVID-19 in patients with multiple myeloma based on a global data network. Blood Cancer J. 2021;11(12):198. Published 2021 Dec 10. doi:10.1038/s41408-021-00588-z.

[13] Brat GA, Weber GM, Gehlenborg N, et al. International electronic health record-derived COVID-19 clinical course profiles: the 4CE consortium. NPJ Digit Med. 2020;3:109. Published 2020 Aug 19. doi:10.1038/s41746-020-00308-0.

[14] Bourgeois FT, Gutiérrez-Sacristán A, Keller MS, et al. International Analysis of Electronic Health Records of Children and Youth Hospitalized With COVID-19 Infection in 6 Countries [published correction appears in JAMA Netw Open. 2021 Jul 1;4(7):e2122388]. JAMA Netw Open. 2021;4(6):e2112596. Published 2021 Jun 1. doi:10.1001/jamanetworkopen.2021.12596.

Challenges of Trustable AI and Added-Value on Health
B. Séroussi et al. (Eds.)

doi:10.3233/SHTI220461

Usability of OMOP Common Data Model for Detailed Lab Microbiology Results

Emmanuelle THERON[a], Jean-François GORSE[a] and Xavier GANSEL[a,1]
[a] *bioMérieux, Data Science, Grenoble, France*

Abstract. Anti-microbial resistance surveillance systems in Europe are limited by the inability to link laboratory data and patient data. The OMOP Common Data Model (OMOP CDM) is an option to store and use patient data in an international context supporting observational research. Detailed medical microbiology laboratory data are usually not stored in OMOP CDM. We propose here a solution to deal with the inherent complexity of microbiology data and store those in the OMOP CDM v5.4. We demonstrate the feasibility of our approach by capturing data from a microbiology in vitro diagnostic middleware, modeling in OMOP CDM 5.4 and querying for visualization.

Keywords. Interoperability, OMOP CDM, data model, microbiology

1. Introduction

VALUE-Dx [1] is a European « Innovative Medicines Initiative » project, that aims to combat antimicrobial resistance (AMR) and improve patient outcome[2]. One of the objectives is to "assess and establish a proof-of-concept data interoperability network to allow connections between laboratory information systems and VALUE-Dx partners". The proliferation of AMR, with the loss of action of antibiotics, has already caused millions of deaths in 2019 [2]. It is estimated that, left unchecked, AMR will cause ~10 million deaths yearly by 2050 [3, 4].

The limits of existing AMR surveillance systems in Europe were highlighted by Tacconelli et al. [5]. This work shows that, among others, purely lab-based systems experience a lack of patient clinical information. Indeed, data management systems and networks, looking to mobilize AMR data, are challenged by the complexity of the data relationships, the practicalities of handling patient identifiable health information, the diversity of data sources, data sovereignty and security issues.

The capabilities of Observational Health Data Science and Informatics (OHDSI) federated network to analyze clinical data from a variety of sources, harmonized in the patient centric interoperable Observational Medical Outcomes Partnership (OMOP) Common Data Model (CDM), are well documented [6]. The need for interoperable laboratory data to support digital medicine and real word data analytics is also documented [7], as well as the mapping of microbiology data across openEHR, Fast

[1] Corresponding Author, Xavier Gansel, bioMerieux, Centre C. Mérieux, 5 rue des Berges, 38000 Grenoble, France; E-mail : xavier.gansel@biomerieux.com

[2] This work has received support from the EU/EFPIA/bioMérieux SA, Janssen Pharmaceutica NV, Accelerate Diagnostics S.L., Abbott, Bio-Rad Laboratories, BD Switzerland Sàrl, and The Wellcome Trust Limited. Innovative Medicines Initiative 2 Joint Undertaking VALUE-Dx grant n° 820755

Healthcare Interoperability Resources (FHIR) and OMOP CDM data standards [8]. Nevertheless, to the best of our knowledge, detailed microbiology laboratory data are not stored in the OMOP CDM for analysis.

To answer some of the unmet needs described by Tacconelli et al. [5], we investigate how detailed clinical microbiology laboratory data from *in vitro* diagnostic systems, can be modeled, incorporated in the OMOP CDM 5.4 [9] and visualized for data analytics needs. This is the first step towards building an AMR data management system able to associate laboratory and patient data.

2. Methods

The OMOP CDM is a relational data model where specific tables are dedicated to the storage of clinical data (e.g. measurement or observation), to the description of care structures (e.g. care_site), or to manage OMOP vocabularies [10]. The main differences between OMOP v5.3 and v6.0 versus v5.4 is the way they handle explicit links between measurements or observations entries and other clinical tables. Where OMOP v5.3 or v6.0 require explicit bidirectional "fact-relationship" association table, the newly published OMOP CDM v5.4 [9] brings, among other features, foreign key usage through modifier_of_event field in measurement table / observation_event_id in the observation table. The original aim is to better support observational cancer research that requires tumors to be characterized by a large set of attributes such as grade, site of origin, biomarkers, etc. [11].

The data used in this paper come from an anonymized dataset containing 996 distinct patients and 4000 microbiology results obtained from bioMérieux® VITEK-MS® and VITEK-2® instruments through the microbiology middleware MYLA®. Alignment between data fields has been carried out using the OHDSI tools WhiteRabbit and RabbitInAHat. We re-used test description mapped to LOINC® and test results (organisms name) mapped to SNOMED CT®[12]. The association between LOINC®, SNOMED CT® codes and OMOP code were realized using ATHENA vocabularies defined by OHDSI. A local OMOP instance was created in a PostgreSQL database and subsequently queried with R, and OHDSI packages SqlRender and DatabaseConnector. Completeness, plausibility and conformance of data loaded in OMOP v5.3 have been assessed using the DataQualityDashboard (DQD) open source tool from OHDSI. Currently, DQD doesn't support OMOP v5.4.

3. Results

3.1. Specific challenges of AMR representation in CDM

Clinical microbiology data is intrinsically complex [13] making modelling in OMOP CDM challenging. One issue is the numerous one to many relationships, indeed one to several pathogens may be identified per specimen and each pathogen may be tested for dozens of antibiotics. Moreover, explicit links between specimen, pathogen identification and AST results need to be retained.

Our analysis revealed that OMOP CDM v5.3 & v6.0 are poorly equipped to model detailed clinical microbiology data. Linking measurements through fact_relationships

(see methods, & figure 1-A) negatively affect the criteria of model interoperability & scalability, query performances and query complexity (due to multiple tables joins).

Leveraging new features of OMOP CDM v5.4 allows explicit links between microbial identification and AST results (see figure 1-B). Our guideline to represent AMR data is presented in figure 1-B. In this representation, a "measurement" is used to capture the lab test and its results (identification or detection). An "observation" captures the name of the identified organism and as many "measurements" as AST tests are linked to the observation (each host both the MIC result & the corresponding category).

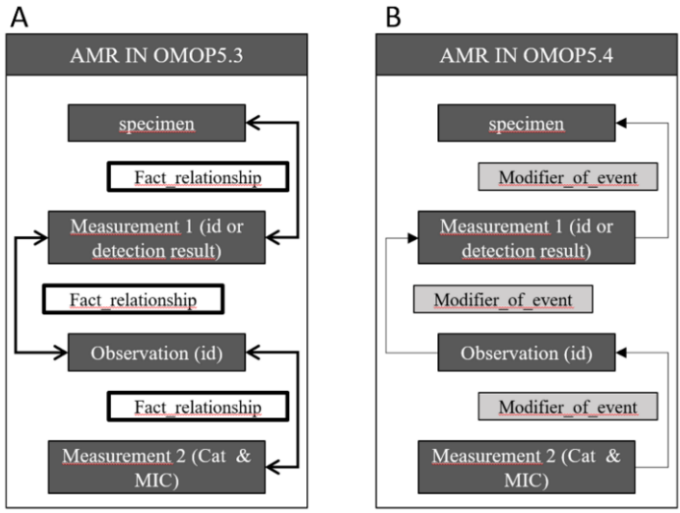


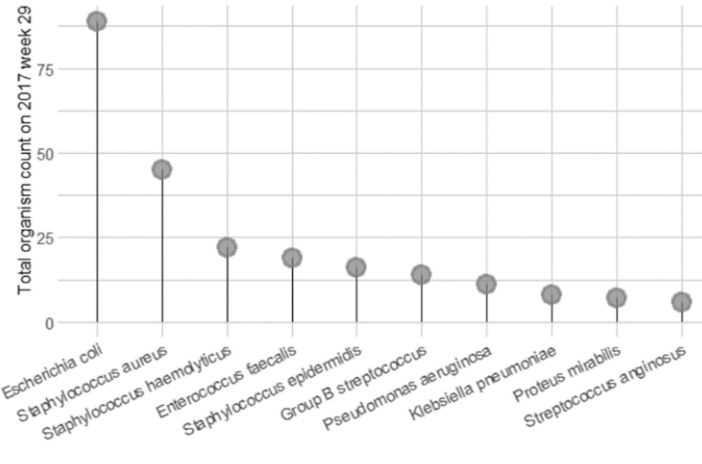
Figure 1: Part A shows modeling in OMOP CDM V5.3 / 6.0 and Part B in OMOP CDM V5.4

3.2. Model evaluation

To assess our model usability and compatibility with OHDSI tools, an implementation of this model has been carried out in a PostgreSQL database with an anonymized dataset (see Methods). Using OHDSI R packages we successfully queried the database and produced visualizations. The example presented in figure 2-A represents the top 10 pathogens identified over a given week in our data set. Data Quality was investigated through 2621 tests carried out with DQD (see figure 2-B). 99% were successful, thus validating the alignment of our model on OMOP standards as well as the plausibility and completeness of our implementation.

The query efficiency formicrobiology data modeled in OMOP CDM 5.4 versus OMOP CDM 5.3 have been compared. A query to obtain the date of a result, the name of the pathogen identified, the name of the antibiotic tested, and the AST status of the organism is five times quicker in our OMOP CDM 5.4 implementation. However, this measure has been performed on a local non optimized setting and will need to be confirmed on a controlled federated architecture.

A



B

	Verification				Validation				Total			
	Pass	Fail	Total	% Pass	Pass	Fail	Total	% Pass	Pass	Fail	Total	% Pass
Plausibility	1893	6	1899	100%	82	0	82	100%	1975	6	1981	100%
Conformance	354	17	371	95%	45	1	46	98%	399	18	417	96%
Completeness	211	6	217	97%	5	1	6	83%	216	7	223	97%
Total	2458	29	2487	99%	132	2	134	99%	2590	31	2621	**99%**

Figure 2: A TOP10 Organisms identified in an example database using OMOP model and OHDSI tools B Summary results of database evaluation using OHDSI DQD tool

4. Discussion

This work not only supports our VALUE-Dx work package objective, but also supports future AMR surveillance networks. Indeed, AMR surveillance is key to informed therapy recommendations, public health interventions, identification of emerging antimicrobial resistance. However, today very few surveillance systems can access data in real-time [5]. Common representation of AMR data across Europe could help surveillance to get up to speed with epidemic process.

In order for our proposed modeling to store every data required for correct AMR surveillance it is paramount that deduplication of the data is achievable. Depending on the organization, the deduplication methods and the data required vary: WHO uses the patient ID, the nature of the sample and the identified pathogen ; whereas the Japanese surveillance system (JANIS) uses the patient ID, the identified pathogen, the AST results and the date of the sample [14]. All this information can be represented in the work presented.

It is also important to be able to differentiate between community-acquired and hospital-acquired infection. This requires the date of admission of the patient to the hospital and more precisely the time between the admission and the collection of the sample. However, this may be poorly represented in a laboratory based system and call for EHR based data, see also below.

Limits of actual surveillance systems have been highlighted by Tacconelli et al. [5] The tendency is for the data to be purely lab-based leading to a lack of patient clinical information. By design OMOP CDM is well equipped to represent clinical outcome, feeding OMOP CDM 5.4 with data originating from both the laboratory and the EHR circumventing these limitations. The modeling proposed here is an option to reach this goal.

5. Conclusions

We propose a model able to represent the complexity of clinical microbiology laboratory data using the newly released OMOP CDM 5.4 [9]. Our work shows that changes in OMOP CDM 5.4, historically pushed by clinical oncology needs [11], benefit clinical microbiology. Compared to OMOP CDM 5.3 it allows for a better data representation and gains in performance.

However additional investigations are needed to confirm those preliminary findings. This should involve a more realistic and larger dataset, additional types of laboratory *in vitro* diagnostics devices, and be carried out in a federated architecture closer to those deployed in clinical settings.

References

[1] VALUE-Dx. VALUE-Dx https://www.value-dx.eu/ [Available from: https://www.value-dx.eu/.
[2] Murray CJL, Ikuta KS, Sharara F, Swetschinski L, Robles Aguilar G, Gray A, et al. Global burden of bacterial antimicrobial resistance in 2019: a systematic analysis. The Lancet. 2022;399(10325):629-55.
[3] Emberger J, Tassone D, Stevens MP, Markley JD. The Current State of Antimicrobial Stewardship: Challenges, Successes, and Future Directions. Current Infectious Disease Reports. 2018;20(9):31.
[4] WHO. Antimicrobial resistance: global report on surveillance. World Health Organization; 2014.
[5] Tacconelli E, Sifakis F, Harbarth S, Schrijver R, van Mourik M, Voss A, et al. Surveillance for control of antimicrobial resistance. Lancet Infect Dis. 2018;18(3):e99-e106.
[6] Maier C, Kapsner LA, Mate S, Prokosch H-U, Kraus S. Patient Cohort Identification on Time Series Data Using the OMOP Common Data Model. Appl Clin Inform. 2021;12(01):057-64.
[7] Gansel X, Mary M, van Belkum A. Semantic data interoperability, digital medicine, and e-health in infectious disease management: a review. Eur J Clin Microbiol Infect Dis. 2019;38(6):1023-34.
[8] Rinaldi E, Thun S. From OpenEHR to FHIR and OMOP Data Model for Microbiology Findings. Stud Health Technol Inform. 2021;281:402-6.
[9] OHDSI. CDM v5.4 https://github.com/OHDSI/CommonDataModel/releases/tag/v5.4.0: OHDSI; 2021 [Available from: https://github.com/OHDSI/CommonDataModel/releases/tag/v5.4.0.
[10] OHDSI. The Book of OHDSI: Observational Health Data Sciences and Informatics; 2021.
[11] Belenkaya R, Gurley MJ, Golozar A, Dymshyts D, Miller RT, Williams AE, et al. Extending the OMOP Common Data Model and Standardized Vocabularies to Support Observational Cancer Research. JCO Clin Cancer Inform. 2021;5:12-20.
[12] Le Gall M, Vachon R, Cellière B, Dante M, Holder F, Mary M, et al. Semantic laboratory data interoperability: Encoding microbiology tests and test results using LOINC ® and SNOMED-CT ® from an IVD manufacturer perspective. In: Microbiology ASf, editor. ASM microbe 2018; 2018 June 10; Atlanta: American Society for Microbiology; 2018.
[13] Fournier P-E, Drancourt M, Colson P, Rolain J-M, Scola BL, Raoult D. Modern clinical microbiology: new challenges and solutions. Nature Reviews Microbiology. 2013;11:574-85.
[14] Kajihara T, Yahara K, Stelling J, Eremin SR, Tornimbene B, Thamlikitkul V, et al. Comparison of de-duplication methods used by WHO Global Antimicrobial Resistance Surveillance System (GLASS) and Japan Nosocomial Infections Surveillance (JANIS) in the surveillance of antimicrobial resistance. PLoS One. 2020;15(6):e0228234.

Challenges of Trustable AI and Added-Value on Health
B. Séroussi et al. (Eds.)

doi:10.3233/SHTI220462

Mapping Korean National Health Insurance Pharmaceutical Claim Codes to SNOMED CT

Hyesil JUNG[a] , Hyeoun-Ae PARK[b,1], Hannah KANG[b], MinSun LEE[b], Sumi SUNG[b] and SeulKi PARK[c]
[a] *Office of eHealth Research and Business, Seoul National University Bundang Hospital, South Korea*
[b] *College of Nursing, Seoul National University, South Korea*
[c] *Department of nursing, Daejeon University, South Korea*

Abstract. The objective of this study was to map pharmaceutical claim codes to SNOMED CT and thereby facilitate multicenter collaborative research and improve semantic interoperability. The claim codes were mapped to SNOMED CT using rule-based automated and manual methods. The maps were internally validated by terminologists and a pharmacist. Finally, 80% of all claim codes were mapped to the concepts of Pharmaceutical/biologic product hierarchy in SNOMED CT. Of them, 50.6% of the codes were exactly mapped to one clinical drug branch concept.

Keywords. Systematized Nomenclature of Medicine Clinical Terms, Semantic interoperability, Pharmaceuticals, National Health Insurance Reimbursement

1. Introduction

Most electronic medical record (EMR) systems in South Korea have been developed in-house and use their own medical vocabulary/terminology to record clinical events or procedures. Thus, the systems cannot exchange data with unambiguous, shared meaning. Although every EMR system uses National Health Insurance (NHI) claim codes [also called electronic data interchange (EDI) codes] to claim a reimbursement in the fee-for-service system, NHI claim codes cannot be used as reference terminology due to limited granularity, non-polyhierarchy, lack of concept identifier (ID) version control, use of a semantic concept ID, a non-unique ID, and a lack of formal definitions [1–2]. Thus, researchers typically map the NHI claim codes to standard terminology in multicenter collaborative studies.

One research team attempted to incorporate the NHI claim codes into Observational Medical Outcomes Partnership (OMOP) vocabulary [2]. However, the pharmaceutical claim codes were not connected to standard terminology such as RxNorm or Systematized Nomenclature of Medicine Clinical Terms (SNOMED CT), limiting their use in clinical research.

[1] Corresponding Author, Hyeoun-Ae PARK, College of Nursing, Seoul National University, 103 Daehak-ro, Jongno-gu, Seoul, 03080, The South Korea; E-mail: hapark@snu.ac.kr

South Korea adopted SNOMED CT as the national standard terminology when the country became the 39th member of SNOMED International in August 2020. SNOMED CT is a globally used reference terminology with extensive coverage: it contains 350,936 active concepts and 1,521,274 descriptions in 19 top-level hierarchies [3]. One of the top-level hierarchies is Pharmaceutical/biologic products with 23,215 active concepts and 170,042 descriptions.

The objective of this study was to map Korea's NHI pharmaceutical claim codes to SNOMED CT and thereby facilitate multicenter collaborative research and improve semantic interoperability.

2. Methods

2.1. Mapping materials

The source data utilized in this study were 22,610 unique pharmaceutical codes covered by the NHI in South Korea published on August 1, 2019. The data included a 9-digit EDI identifier (product ID), the Korean brand name, the manufacturer, the WHO Anatomical Therapeutic Chemical Classification (ATC) code, and the ATC name. Missing information for pharmaceuticals that could not be identified from the EDI code was obtained from the FirstDIS Ltd. database, which is a repository containing information related to the pharmaceuticals used in South Korea.

The target data were restricted to the concepts in 'Pharmaceutical/biologic products', 'Substance', 'Organism', or 'Physical object' top-level hierarchies in the SNOMED CT international edition released on July 31, 2019. If no appropriate concepts appeared in the international edition, we used the Argentinian and United States extensions released on November 30, 2019.

2.2. Mapping methods

Figure 1 presents the process used to map pharmaceutical claim codes to SNOMED CT concepts and validate the results. Creating the maps was an iterative process of rule development and application to search for (map) the optimal target concept.

A rule-based mapping system was developed utilizing Snowstorm, which provides the terminology server API for the SNOMED international browser, including the international edition and about 14 local extensions. Snowstorm is built on top of Elasticsearch and applies the term-based search using multiple prefix, any order, matching. The first part of one or more words input can be used to match the descriptions of concepts. If a description term does not exactly match all the prefixes in the search term, it will be excluded from the results.

When the mapping system provides candidates that match the search terms, the first author reviewed the candidates and selected the optimal SNOMED CT concept among them. If no suitable concept was found among the candidates, she searched for the appropriate concept and mapped them manually. Pharmaceuticals consisting of three or more ingredients were also manually mapped to SNOMED CT.

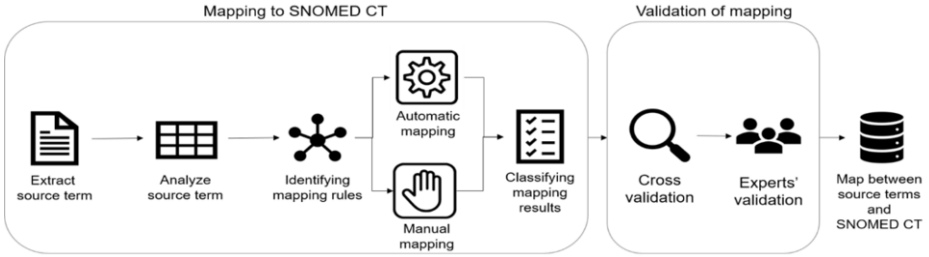


Figure 1. Research process

2.2.1. Creating the maps

2.2.1.1. Developing rules for mapping

We analyzed the source data and extracted the ingredients (base ingredient and active ingredient), strength (numerator value and numerator units), dose form, and route of administration information into a structured data format.

We randomly selected 600 source data and manually mapped them to SNOMED CT by combining the pharmaceutical information extracted in the previous step to develop rules for automatic mapping. First, we combined ingredient(s), strength, dose form, and administration route and searched whether the relevant SNOMED CT concept existed in the clinical drug branch of the Pharmaceutical/biologic product hierarchy. If the relevant concept did not exist in the clinical drug branch, we only combined ingredient(s) and administration route to search for the concept in the medicinal product form branch. Figure 2 presents the flow of combining structured elements (information) from the source data and mapping to the relevant concepts within the SNOMED CT hierarchy.

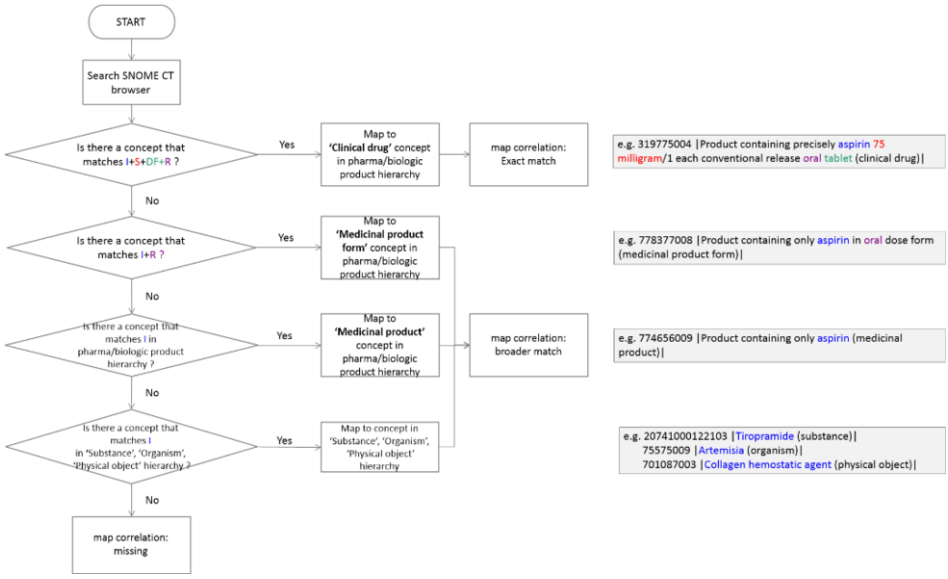


Figure 2. Mapping rules for combining structured elements (information) from source data
I-ingredient; S-strength; DF-dose form; R-route of administration

2.2.1.2. Search strategies

Three types of drug-specific normalization rules [4-5] for managing the variation in clinical drug names were applied to search for appropriate target concept; (1) expanding abbreviated words (2) reformatting specific parts of the drug name (3) removing salt modifiers in ingredient names.

For example, shortened dose forms such as 'tab', 'susp', and 'cap' were expanded to 'tablet', 'suspension' and 'capsule', respectively. In terms of reformatting, 'syrup' and 'cutaneous emulsion' were normalized to 'oral suspension' and 'cutaneous lotion', respectively. 'Oral suspension' and 'oral liquid' were changed to 'oral solution'. And when an ingredient's strength value was less than 1, we converted the strength unit (e.g. 0.1 g →100 mg). The ingredient names of clinical drugs sometimes have the salt modifiers and sometimes do not. We removed the salt modifiers from the ingredient name when there was no appropriate target concept with the salt modifiers.

The map correlations were classified as 'exact' match, 'broader' match, and 'missing' according to the extent of coverage of the target concept compared to the source code [5-6]. If a target concept was consistent with the source code in terms of ingredients, strength, dose form, and administration route, the source code had the 'exact' match. If a target concept partially matched the source code, the source code had a 'broader' match. If there was no ingredient for the source code in the target concept, the source code had 'missing' as the map correlation.

Map cardinality was defined as '1:1'and '1: N', according to the number of target concepts mapped to a single source code.

2.2.2. Validation

The maps were validated by another mapper with expert experience in SNOMED CT mapping. When the two mappers agreed on the mapping results, the results were considered internally valid. If they disagreed, the results were discussed in group meetings attended by the project manager and other researchers who were not involved in the mapping process. Any concepts not agreed upon during the group discussion or with the 'missing' correlations were validated by a pharmacist with experience in SNOMED CT and Identification of Medicinal Product mapping experience.

The final maps consisted of a source code, the map correlation type, the SNOMED CT concept identifier, fully specified name, hierarchy, and pre-coordinated expression.

3. Results

There was good agreement (79.9%) between mappers in the process of mapping pharmaceutical claim codes to SNOMED CT concepts. A pharmacist externally validated 250 concepts that were not agreed upon throughout the group discussion.

Of the 22,610 source data, almost 80% were mapped to the concepts of Pharmaceutical/biologic product hierarchy, and 11,453 (50.6%) of these codes were exactly mapped to one (1:1 map) 'clinical drug' branch concept. In total, 11,151 (49.3%) of the claim codes mapped to a broader concept(s), and 1233 of these codes were mapped to two or more (1:N map) target concepts. However, five pharmaceuticals, including Revanex, Noltec, Ulistin, and Godex did not map to any target concept.

4. Discussion

To develop mapping rules, we analyzed the source codes and extracted pharmaceutical information including the ingredients, strength, dose form, and administration route. First, all information (elements) was combined to search for the relevant target concept. If there was no relevant concept, elements were gradually omitted and mapped to a broader target concept. Almost 50% of the source codes were mapped to broader concepts. In particular, multi-ingredient pharmaceuticals with strength specified for each ingredient were challenging [5]. When conducting research in which the strength and dose form of pharmaceuticals are important, broad concepts make it impossible to provide information on strength and dose form, which may lower the reliability and accuracy of the study. As mentioned by Zhou et al., one possible solution is post-coordinate two or more SNOMED CT concepts that are semantically representative of a single claim code [5].

The five claim codes that were not mapped to any SNOMED CT concept were for pharmaceuticals used only in South Korea or certain countries such as Japan, China, or India. They were pharmaceuticals with ilaprazole, revaprazan, and ulinastatin as ingredients. The Korean extension to be developed soon may include these concepts.

5. Conclusions

The results of mapping Korea's NHI pharmaceutical claim codes to SNOMED CT concepts will facilitate semantic interoperability and collaborative research. The mapping rules and methodologies developed in this study can be used to map between expired or newly created pharmaceutical claim codes and SNOMED CT concepts.

Acknowledgments

This research was supported by the Korea Health Information Service and a National Research Foundation of Korea grant funded by the Ministry of Science and ICT (no. NRF-2021R1A2C1091261).

References

[1] Cimino JJ. Desiderata for controlled medical vocabularies in the twenty-first century. Methods Inf Med. 1998 Nov;37(4-5):394–403.

[2] Seong Y, You SC, Ostropolets A, Rho Y, Park J, Cho J, Dymshyts D, Reich CG, Heo Y, Park RW. Incorporation of Korean electronic data interchange vocabulary into observational medical outcomes partnership vocabulary. Healthc Inform Res. 2021 Jan;27(1):29–38.

[3] SNOMED International [Internet]. General Release Statistics. Available from: https://browser.ihtsdotools.org/qa/#/generalReleaseStatistics

[4] Peters L, Kapusnik-Uner JE, Bodenreider O. Methods for managing variation in clinical drug names. AMIA Annu Symp Proc. 2010 Nov;13:637-41.

[5] Zhou L, Plasek JM, Mahoney LM, Chang FY, DiMaggio D, Rocha RA. Mapping Partners Master Drug Dictionary to RxNorm using an NLP-based approach. J Biomed Inform. 2012 Aug;45(4):626-33.

[6] Merabti T, Massari P, Joubert M, Sadou E, Lecroq T, Abdoune H, Rodrigues JM, Darmoni SJ. An automated approach to map a French terminology to UMLS. Stud Healht Technol Inform. 2010;160:1040-4.

Challenges of Trustable AI and Added-Value on Health
B. Séroussi et al. (Eds.)

doi:10.3233/SHTI220463

OntoBioStat: Supporting Causal Diagram Design and Analysis

Thibaut PRESSAT LAFFOUILHÈRE[a,b,c,1], Julien GROSJEAN[a,d], Jacques BÉNICHOU[b,e], Stefan J. DARMONI[a,d] and Lina F. SOUALMIA[c,d]
[a] *CHU Rouen, Department of Biomedical Informatics, F-76000 Rouen, France*
[b] *CHU Rouen, Department of Biostatistics, F-76000 Rouen, France*
[c] *Normandie Univ, UNIROUEN, LITIS-TIBS EA 4108, F-76000 Rouen, France*
[d] *LIMICS U1142, Sorbonne Université, Paris, France*
[e] *INSERM U1018, CESP, Université Paris-Saclay, Paris, France*

Abstract. Suitable causal inference in biostatistics can be best achieved by knowledge representation thanks to causal diagrams or directed acyclic graphs. However, necessary and sufficient causes are not easily represented. Since existing ontologies do not fill this gap, we designed OntoBioStat in order to enable covariate selection support based on causal relation representations. OntoBioStat automatic ontological causal diagram construction and inferences are detailed in this study. OntoBioStat inferences are allowed by Semantic Web Rule Language rules and axioms. First, statements made by the users include outcome, exposure, covariate, and causal relation specification. Then, reasoning enable automatic construction using generic instances of Meta_Variable and Necessary_Variable classes. Finally, inferred classes highlighted potential bias such as confounder-like. Ontological causal diagram built with OntoBioStat was compared to a standard causal diagram (without OntoBioStat) in a theoretical study. It was found that confounding and bias were not completely identified by the standard causal diagram, and erroneous covariate sets were provided. Further research is needed in order to make OntoBioStat more usable.

Keywords. Causality, Ontology, Statistic, Bias, Variable selection, Decision support techniques

1. Introduction

According to the Medical Subject Heading (MeSH) thesaurus, '[…] Causes are termed **necessary** when they must always precede an effect and **sufficient** when they initiate or produce an effect [...]'.

In causal inference the aim of the statistical analysis is to provide an unbiased causal effect of an exposure of interest on an outcome (e.g., effect of an oral antidiabetic on pancreatic cancer risk), using for example adjustment methods [1]. Causal diagrams are used in order to select the right sets of covariates that should be adjusted for [2]. Causal diagrams (CDs) are qualitative representations of a given study, with variables as nodes and probabilistic causal relations as edges between variables. CD's representation depends on the use case and there is no tutorial or universal rules

[1] Corresponding Author, Thibaut PRESSAT LAFFOUILHÈRE, 37 Boulevard Gambetta, 76000 Rouen, France; E-mail: t.pressat@chu-rouen.fr.

that could help users to build a causal diagram with necessary and sufficient causes in all cases [3,4]. Existing published ontologies such as the Relation Ontology [5] or Radiology Gamuts Ontology [6] do not cover entirely the complexity of the different causal relations needed for covariate selection (i.e. distinction between counterfactual probabilistic, sufficient and necessary causes). We designed the OntoBioStat [7] ontology in order to support covariate selection for causal inference. OntoBioStat was built using expert knowledge corpus, theoretical cases, and literature review in order to address several competency questions. OntoBioStat is a domain ontology that can help users in their tasks of building and understanding causal diagrams.

This paper focuses on two OntoBioStat features: (i) automatic construction of causal diagrams with necessary causes (called in this article ontological causal diagram), and (ii) reasoning on necessary and sufficient causes. It is divided in two parts: first the description of the classes, relations, rules and instances involved in each of the two features, and then a theoretical study relying on necessary and sufficient causes was presented.

2. Materials and Methods

2.1. OntoBioStat

OntoBioStat was built with the Protégé software [8]. The last version is available at https://bioportal.bioontology.org/ontologies/OBS. Reasoning is supported by the Pellet reasoner [9]. OntoBioStat is composed by 53 classes and 33 relations. Here we focused on 24 classes, five object properties, no data properties, 28 instances, and nine rules from OntoBioStat. Indeed, OntoBioStat knowledge representation includes for example interaction and missing data that are not relevant here (inferences are not impacted). The following classes were used: “Exposure_stressor”, “Outcome” and “Covariates”. “Meta_Variable” class groups “Decision-makers” and “Method_of_measurement”. “Theoretical_Variable” class groups “Condition”, “Environment”, “Status”, “Health_Behavior”, “Intervention_Effect”. “Necessary_Variable” class groups “Exist”, “Available”, “Indicated”, “Prescribed”, “Delivered”, “Investigated”, and “Adhered”. Inferred classes presented were “Reverse_Causality” (subClassof “Outcome”), and “Unadjusted_Confounder” (both bias that cannot be corrected using adjustment), “Mediation_Differential_Confounder” and “Confounder-like” (subClassof “Covariate”) (both bias that should be corrected using adjustment). Object properties *Related_to* and his descendants were used to represent unidirectional, bidirectional, and non-directional probabilistic ‘causal’ relations between two instances. Among *Signed* properties, *Contraindication* and *Absolute_Indication* are the two object properties that represent sufficient (deterministic) causal relations. They were named with ‘indication’ term because most of the time the sufficient causes in biomedical research are patients characteristics that contraindicate or impose the use of a particular treatment.

“Necessary_Variable” and “Meta_Variable” were fed with 28 generic instances that could be used for any study. These instances are not involved in any causal relation until first new instances are added and the reasoner activated. Automatic ontological causal diagram (OCD) construction relies on five Semantic Web Rule Language (SWRL) rules. The inferred classes rely on three rules for necessary causes (see example below (1)) and one axiom for the sufficient causes. Several SWRL rules about

causal reasoning were used to infer all *Related_to* descendants based on *isCauseof* statements that are not developed here.

Inverse_Directed_Relation(?x,?y)^Mediator(?y)^Necessary_Variable(?x) -> Mediation_Differential_Confounder (1)

Figure 1. Decision-support pipelines and steps (1: User enters the two first variables names as instances and define their classes: (i) Exposure_stressor and Outcome, (ii) Theoretical_Class, 2: User activates reasoner, 3: Automatic construction using existent generic instances (purple) and causal relations (black), 4: User adds covariates and causal relations (red), 5: User refreshes the reasoner, 6: OntoBioStat provide inferred classes (yellow) and new object properties (blue), 7: Lecture of the results and inferences explanations.)

Decision-support based on OntoBioStat requires several exchanges of information between the biostatistician and the Protégé (Figure 1).

2.2. Theoretical study

The aim was to obtain an unbiased true causal effect between the use of oral antidiabetics (oad 1 versus oad 2) and time to pancreatic neoplasm diagnosis. Covariates included comorbidities, oad side effects, and co treatment. Based on the following statements, an OCD and a CD were built: (i) patient comorbidity contraindicates the use of oad 1 and may cause a pancreatic neoplasm, (ii) co treatment is prescribed with oad 2 and is known to cause pancreatic neoplasm, (iii) side effects more often caused by oad 2 more often lead to medical consultation.

The OCD created and analyzed with OntoBioStat was confronted to a CD. The CD was created without necessary variables provides by OntoBioStat during automatic construction process. The CD reasoning to solve variable selection was based on the back-door criterion algorithm [10] instead of rules and axioms from OntoBioStat. Reverse causality implies a cyclic graph; hence the directed causal relation from Outcome to Exposure was not included in the CD.

3. Results

For more readability, OCD inferences are presented in one truncated diagram excluding some of the necessary variables and inferred object properties (Figure 2). Explanations about the inferences are the following: (i) ‘co treatment’ is a “Confounder-like” variable because ‘co treatment’ *isCauseof* “Outcome” and *hasCause* ‘prescribed_exposure’ that is an “Indirect_Confounder”; (ii) ‘side effects’ is a “Mediation Differential Confounder” because *hasCause* “Exposure” and *isCauseof* “Necessary_Variable” (1); (iii) ‘comorbidity’ is an *Unadjusted Confounder* because *Contraindication* of “Exposure” and *isCauseof* “Outcome”; (iv) “Outcome” is

classified in "Reverse_Causality" because "Outcome" *isIndirectCauseof* "Exposure". The standard CD is represented with inferences in Figure 3. Without necessary variable 'co treatment' and 'side effects' are seen as covariates that do not bias the true causal effect, but as covariates that must not be selected for adjustment (mediator) which may increase bias. Even with necessary variables specification 'side effects' covariate requires adequate reasoning to be considered as a potential candidate for adjustment. Without sufficient cause specification, the covariate 'comorbidity' is seen as a potential candidate for adjustment whereas this adjustment cannot correct bias.

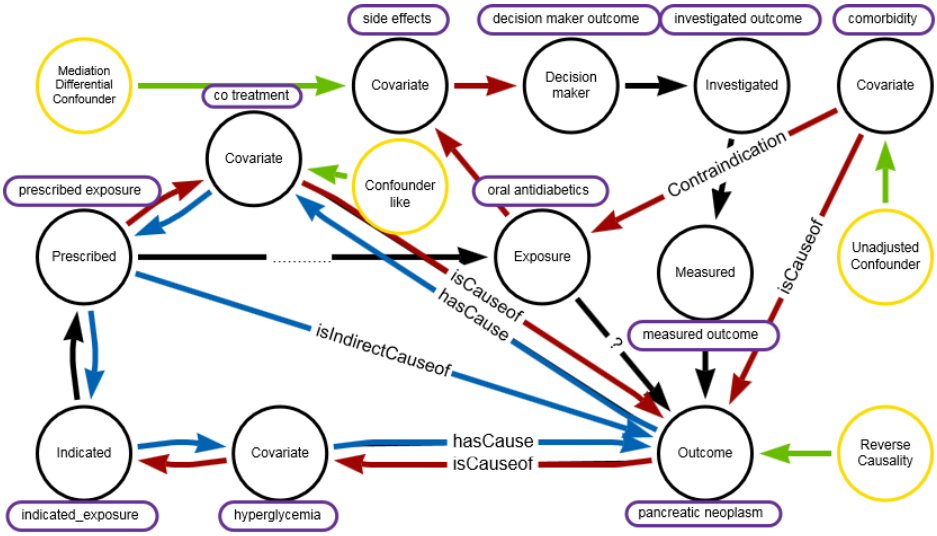


Figure 2. Ontological causal diagram built with OntoBioStat using inferences. Instances name are in purple. Object properties stated are in red, inferred in blue, being a part of automatic construction black. Inferred classes are in yellow.

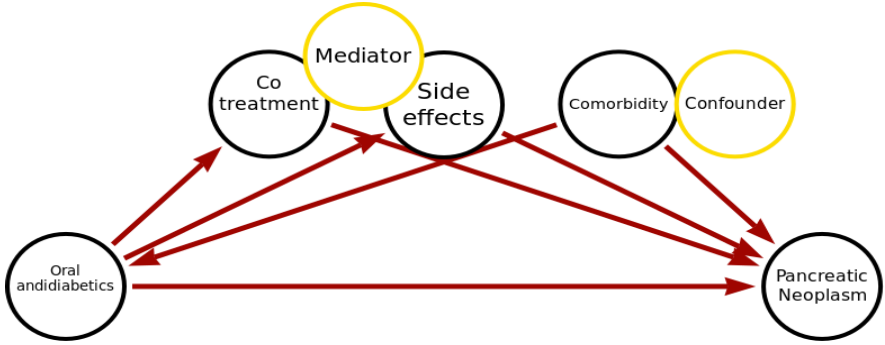


Figure 3. Standard causal diagram with inferences.

4. Discussion and Conclusions

In this article, we showed the usefulness of a novel model following the footsteps of the Directed Acyclic Graph (DAG), CD, and Sufficient Component Cause model [11] that could be used to enhance the consciousness of the study biases. Furthermore, the

pre-existent generic instances provide a significant added value to the knowledge representation of a given study and should help users to reflect on their own practices. Since the aim of OntoBioStat is to support covariate selection, it relies on a sufficient formalism. For example, it does not include distinction between state, event and process or between 'allow', 'maintain', 'perpetuate' relations as defined in the ontology of causal relations [12].

Decision-support systems such as dagitty [13,14] for DAGs provide an easy to use interface. The R package dagitty enables users to specify CD's structure and to obtain the right set of covariates (minimal and sufficient), instrumental variable, and path analysis. However, dagitty does not provide an automatic DAG construction, adapted reasoning for necessary or sufficient cause, nor rich explanation of the results. Actually, OntoBioStat may be seen more as an educational tool or a safety net provider for unskilled biostatistics users than a real decision-support system to be used on daily basis by expert users for two main reasons: (i) reasoning based on rules and axioms do not provide minimal sufficient set of covariates but put forward all covariates that could bias the results, hence minimal set have to be selected manually, (ii) biostatisticians are not familiar with the Protégé.

Directions for future research include: (i) the implementation of OntoBioStat as an operational system named MetBRaYN [7], combining the strengths of dagitty and OBS with an R interface; (ii) the mapping of ontologies object properties with OntoBioStat causal object properties in order to feed with several instances the ontology.

References

[1] Grimes DA, Schulz KF. Bias and causal associations in observational research. The Lancet. 2002 Jan;359(9302):248–52.
[2] Greenland S, Pearl J, Robins JM. Causal diagrams for epidemiologic research. Epidemiology. 1999 Jan;10(1):37–48.
[3] Vanderweele TJ, Robins JM. Empirical and counterfactual conditions for sufficient cause interactions. Biometrika. 2008 Jan 31;95(1):49–61.
[4] Digitale JC, Martin JN, Glymour MM. Tutorial on directed acyclic graphs. Journal of Clinical Epidemiology. 2021 Aug;S0895435621002407.
[5] Smith B, Ceusters W, et al. Relations in biomedical ontologies. Genome Biol. 2005;6(5):R46.
[6] Kahn CE. Transitive closure of subsumption and causal relations in a large ontology of radiological diagnosis. Journal of Biomedical Informatics. 2016 Jun;61:27–33.
[7] Pressat Laffouilhère T, Grosjean J, et al. Ontological Models Supporting Covariates Selection in Observational Studies. Stud Health Technol Inform. 2021 May 27;281:1095–6.
[8] Musen, M.A. The Protégé project: A look back and a look forward. AI Matters. Association of Computing Machinery Specific Interest Group in Artificial Intelligence, 1(4), 2015 Jun.
[9] Sirin E, Parsia B, et al. Pellet: A practical OWL-DL reasoner. Web Semantics. 2007 Jun;5(2):51-53.
[10] Pearl J. Causality: models, reasoning, and inference. Cambridge, U.K. ; New York: Cambridge University Press; 2000. 384 p.
[11] Rothman KJ, Greenland S. Causation and Causal Inference in Epidemiology. Am J Public Health. 2005 Jul;95(S1):S144–50.
[12] Galton A. States, Processes and Events, and the Ontology of Causal Relations. Frontiers in Artificial Intelligence and Applications. Volume 239: Formal Ontology in Information Systems. 279–292.
[13] Textor J, van der Zander B, Gilthorpe MS, Liśkiewicz M, Ellison GTH. Robust causal inference using directed acyclic graphs: the R package 'dagitty'. Int J Epidemiol. 2017 Jan 15;dyw341.
[14] Ankan A, Wortel IMN, Textor J. Testing Graphical Causal Models Using the R Package "dagitty". Current Protocols. 2021 Feb.

Challenges of Trustable AI and Added-Value on Health
B. Séroussi et al. (Eds.)

doi:10.3233/SHTI220464

Mapping of ICD-O Tuples to OncoTree Codes Using SNOMED CT Post-Coordination

Tessa OHLSEN[a,1], Valerie KRUSE[b], Rosemarie KRUPAR[c], Alexandra BANACH[a], Josef INGENERF[a,d] and Cora DRENKHAHN[d]
[a] *Institute for Medical Informatics, University of Lübeck, Lübeck, Germany*
[b] *Clinic for Hematology and Oncology, UKSH, Lübeck, Germany*
[c] *Pathology of the Research Center Borstel, Leibniz Lung Center, Borstel, Germany*
[d] *IT Center for Clinical Research, University of Lübeck, Lübeck, Germany*

Abstract. Around 500,000 oncological diseases are diagnosed in Germany every year which are documented using the International Classification of Diseases for Oncology (ICD-O). Apart from this, another classification for oncology, OncoTree, is often used for the integration of new research findings in oncology. For this purpose, a semi-automatic mapping of ICD-O tuples to OncoTree codes was developed. The implementation uses a FHIR terminology server, pre-coordinated or post-coordinated SNOMED CT expressions, and subsumption testing. Various validations have been applied. The results were compared with reference data of scientific papers and manually evaluated by a senior pathologist, confirming the applicability of SNOMED CT in general and its post-coordinated expressions in particular as a viable intermediate mapping step. Resulting in an agreement of 84,00 % between the newly developed approach and the manual mapping, it becomes obvious that the present approach has the potential to be used in everyday medical practice.

Keywords. ICD-O, OncoTree, SNOMED CT, Ontoserver, HL7 FHIR, post-coordination, terminology server

1. Introduction

In a Molecular Tumor Board (MTB), an interdisciplinary team of physicians creates therapy recommendations for patients with oncological diseases beyond standard treatment options. In the MTB of the University Medical Center Schleswig-Holstein (UKSH), the software cBioPortal [1] shall be used to visualize molecular genetics and clinical data and to support decision-making. cBioPortal uses OncoTree, a hierarchically organized structure for the classification of currently 868 tumor types [2]. By considering a tumor's histology and localization, it can be matched to a node of the OncoTree.

In pathology reports, neoplasms are routinely coded using the ICD-O classification [3]. ICD-O differentiates between codes of two axes – topography and morphology – which are combined into a tuple. While a mapping from OncoTree to ICD-O is available

[1] Corresponding Author, Tessa Ohlsen, Institute of Medical Informatics, University of Lübeck, Ratzeburger Allee 160, 23564 Lübeck, Germany; E-mail: tessa.ohlsen@student.uni-luebeck.de

[2], the given associations are inherently unidirectional due to ICD-O being far more granular. Although ICD-O and OncoTree are both classifications for the oncology domain, individual class boundaries are not necessarily similar and thus class extensions only overlap partly. Due to this discrepancy, a simple translation from ICD-O to OncoTree is troublesome so that a more sophisticated approach is needed. For our project, we hypothesized that SNOMED CT may work as a feasible intermediate mapping step. To additionally minimize the workload and sources of error during the mapping process, a multi-step procedure was developed to map ICD-O tuples to OncoTree codes in a preferably (semi-)automatic way.

2. Methods

Four steps were developed to map the ICD-O tuples to OncoTree. SNOMED CT, as the most comprehensive terminology in medicine [4], was used as a purpose-agnostic intermediate representation to mediate between ICD-O and OncoTree. Figure 1 shows an overview of the individual mapping steps, which are explained in the following. All steps of the mapping were implemented using a local instance of the HL7 FHIR (Fast Healthcare Interoperability Resources)-based terminology server Ontoserver [5].

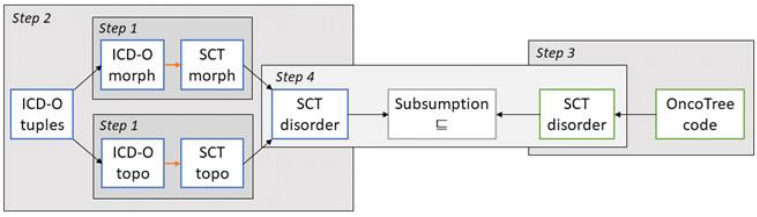


Figure 1. Sequence of mapping ICD-O tuples to the OncoTree with the four steps: (1) Automapping, (2) mapping of ICD-O tuples to SNOMED CT (SCT) disorder codes, (3) mapping of OncoTree codes to SNOMED CT disorder codes, and (4) subsumption testing between SCT-mapped ICD-O tuples and SCT-mapped OncoTree codes.

2.1. Automapping

In this preliminary mapping step, the two axes of ICD-O tuples (topography and morphology) are considered independently. All existing topography and morphology codes are provided by the World Health Organization (WHO), totaling up to 327 topography and 1090 morphology codes [3].

Each of the ICD-O topography and morphology codes shall be converted into SNOMED CT concepts. To achieve this, an Ontoserver-associated web application called Snapper[2] can be employed. Despite being primarily an editing tool for FHIR-based terminology resources, Snapper also offers an automapping feature, which has proven to be reliable and efficient when mapping to SNOMED CT [6]. So, the ICD-O codes are imported into Snapper and automatic mapping suggestions generated with appropriate settings: The target range for ICD-O topography and morphology codes can be limited to SNOMED CT concepts of the subhierarchies of body structures and morphologic abnormalities, respectively.

[2] https://ontoserver.csiro.au/snapper2/

Afterwards, a manual post-processing (choosing from suggestions or augmenting them via manual search) is completed by the PhD student. The resulting relations between ICD-O topography or morphological codes and the corresponding SNOMED CT concepts are stored in separate FHIR ConceptMaps [7] on the Ontoserver.

2.2. Mapping of ICD-O tuples to SNOMED CT disorder codes

Based on the previous step, the mapping of combined ICD-O tuples is performed. Here, the input dataset consists of 1800 ICD-O tuples used in tumor documentation at UKSH, Campus Lübeck since 2016. For each tuple, the two SNOMED CT codes corresponding to its topography and morphology can be used as the basis for the semi-automated detection of a pre-coordinated concept or the automated generation of a post-coordinated expression. In both cases, *64572001 | Disease |* is used as the central "focus" concept which is further refined according to the SNOMED CT Concept Model via the attributes *363698007 |Finding site|* and *116676008 |Associated morphology|* with the respective body structure and morphologic abnormality concepts. Like before, mapping results are collated into two separate ConceptMaps.

Pre-coordination

The pre-coordinated approach uses the SNOMED CT Expression Constraint Language (ECL) to find predefined concepts which fulfill the given expression. According to the basic structure described above, an example ECL expression is as follows:

< 64572001 | Disease | :
{ 363698007 | Finding site | = 39607008 | Lung structure | ,
116676008 | Associated morphology | = >! 35917007 | Adenocarcinoma | }

This expression queries all diseases which are found at the lungs with a morphology of adenocarcinoma or one of its direct parent concepts and would yield *707451005 |Primary adenocarcinoma of lung|* as a result.

An iterative algorithm was developed for retrieving a fitting pre-coordinated concept. In the first iteration, it considers only the exact attribute-value pairs as defined above. In subsequent iterations, further levels of parent concepts are considered for the topography and/or morphology concept, making the expression increasingly more general. The algorithm terminates as soon as at least one result is found or after a maximum of 14 iterations. If multiple results are returned, the best option is chosen interactively.

Post-coordination

The second approach makes use of SNOMED CT Postcoordinated Expressions (PCE) which allow for the flexible combination of multiple concepts into previously unrepresentable meanings. Thus, a granularity beyond the scope of predefined concepts is conceivable. While PCEs look similar to ECL expressions on a syntactic level, they represent valid SNOMED CT "concepts" instead of queries. So, each ICD-O tuple is mapped to a PCE constructed according to the partwise mapping results and the basic structure described above, e.g.:

64572001 | Disease | :
{ 363698007 | Finding site | = 39607008 | Lung structure | ,
116676008 | Associated morphology | = 35917007 | Adenocarcinoma | }

2.3. Mapping of OncoTree codes to SNOMED CT disorder codes

The 868 OncoTree codes are converted manually to SNOMED CT disorder codes by the PhD medical student. Here again, post-coordinated SNOMED CT expressions are built following the known structure and mapping results stored in a separate ConceptMap on the Ontoserver.

2.4. Subsumption testing and overall mapping

After mapping both the ICD-O tuples and OncoTree codes to SNOMED CT disorder codes, subsumption relations between both can be identified pairwise by querying the Ontoserver with the HL7 FHIR operation "$subsumes". Only the results "equivalent" and "subsumes" imply a relevant association from the respective ICD-O tuple to the OncoTree code which is then stored in yet another ConceptMap for the overall mapping.

3. Results

From the previously mentioned dataset, 105 of 1800 ICD-O tuples were identified as invalid and had to be excluded. Of the remaining 1695 ICD-O tuples used, 99.23 % could be successfully mapped to an OncoTree code using the pre-coordinated approach. With post-coordination, a mapping could be achieved for all input tuples. For 63.24 % of mapping relations, the selected target OncoTree codes are equivalent between both approaches with the code chosen via post-coordination being more specific and thus more precise otherwise.

The most frequent 100 ICD-O tuples already cover 63.00 % of all oncological diseases registered at UKSH, Campus Lübeck since 2016. A senior pathologist previously not involved in the process manually mapped this excerpt as reference for determining mapping accuracy. 84.00 % were found to be equivalent with the post-coordinated approach and 56.00 % using pre-coordination. For another validation, a previously published mapping by Thomas et al. [8] was used as reference data. 77.92 % of results based on the superior post-coordinated approach matched with the reference data. Otherwise, the reference data were 1.48 levels deeper in the OncoTree on average.

4. Discussion

By implementing a multi-step process, a semi-automated mapping from ICD-O tuples to OncoTree codes could be achieved successfully and with decent accuracy. SNOMED CT was found to be a workable solution to both bridge the gap between disparate classifications and to support automatization, especially attained by employing advanced features like ECL, post-coordination, and subsumption testing. Utilizing post-coordinated expressions also proved useful in achieving more precise mapping results by covering a broader scope as well as by preventing the loss of information inevitable when limited to the predefined combinations of pre-coordinated concepts.

A pre-requisite for implementing the described mapping approach was the availability of appropriate tooling. Here, the FHIR-based Ontoserver convincingly

supported the standardized access of terminology content and related operations, including the previously described specialized SNOMED CT features.

Nevertheless, validation revealed some inaccuracies in the mapping results which can be mainly attributed to three issues. Firstly, during validation, only a binary measure for equivalence was utilized. But, despite being not exactly the same, many results are still semantically similar. Secondly, OncoTree currently provides - with only 868 classes - significantly less detail compared to ICD-O and SNOMED CT, making the mapping inherently imprecise and non-reversible. Thirdly, further discrepancies between the involved terminologies hinder the mapping process. E.g., OncoTree's structure follows pragmatic considerations of everyday clinical practice which are sometimes incompatible with SNOMED CT's strictly logical polyhierarchy.

To mitigate some of these issues, further evaluations considering the semantic distance between divergent results and the specific influence of using SNOMED CT as an intermediate representation are in progress.

5. Conclusion

A largely automated mapping of ICD-O tuples to OncoTree codes could be implemented successfully by using SNOMED CT as an intermediate step. SNOMED CT, in combination with HL7 FHIR operations and a terminology server, enables a straightforward implementation. The approach using post-coordination outperformed the pre-coordination variant both in mapping coverage and accuracy. The results can easily be expanded to further ICD-O tuples and will be integrated into cBioPortal in the future.

Acknowledgements

This work was supported by the HiGHmed project within the Medical Informatics Initiative, funded by the Federal Ministry of Education and Research (grant number 01ZZ1802Z).

References

[1] Gao J, Aksoy BA, Dogrusoz U, Dresdner G, Gross B, Sumer SO, et al. Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. Sci Signal. 2013 Apr 2;6(269):pl1.

[2] Kundra R, Zhang H, Sheridan R, Sirintrapun SJ, Wang A, Ochoa A, et al. OncoTree: A Cancer Classification System for Precision Oncology. JCO Clinical Cancer Informatics. 2021 Mar;(5):221–30.

[3] Staff WHO, Organization WH, Jack A, Percy C, Sobin L, Whelan S. International Classification of Diseases for Oncology: ICD-O. World Health Organization; 2000. 252 p.

[4] SNOMED - 5-Step Briefing [Internet]. SNOMED. [cited 2022 Jan 20]. Available from: https://www.snomed.org/snomed-ct/five-step-briefing

[5] Metke-Jimenez A, Steel J, Hansen D, Lawley M. Ontoserver: a syndicated terminology server. J Biomed Semantics. 2018 Sep 17;9(1):24.

[6] Drenkhahn C, Burmester S, Ballout S, Ulrich H, Wiedekopf J, Ingenerf J. Using FHIR terminology services-based tools for mapping of local microbiological pathogen terms to SNOMED CT at a German university hospital. In 2020. Available from: https://dvmd.de/wp-content/uploads/2020/11/A-148.pdf

[7] ConceptMap - FHIR v4.0.1 [Internet]. [cited 2022 Jan 21]. Available from: https://www.hl7.org/fhir/conceptmap.html

[8] Thomas S, Lichtenberg T, Dang K, Fitzsimons M, Grossman RL, Kundra R, et al. Linked Entity Attribute Pair (LEAP): A Harmonization Framework for Data Pooling. JCO Clinical Cancer Informatics. 2020 Nov 1;(4):691–9.

Challenges of Trustable AI and Added-Value on Health
B. Séroussi et al. (Eds.)

doi:10.3233/SHTI220465

Indexing Imaging Reports for Data Sharing: A Study of Mapping Using RadLex Playbook and LOINC

Pierre LEMORDANT[a,b,1], Fleur MOUGIN[c], Sandie CABON[a], Yves GANDON[d], Guillaume BOUZILLÉ[a] and Marc CUGGIA[a]
[a] *Univ Rennes, CHU Rennes, Inserm, LTSI - UMR 1099, F-35000 Rennes, France*
[b] *Enovacom, Marseille, France*
[c] *Univ. Bordeaux, INSERM, BPH, U1219, Bordeaux, France*
[d] *CHU Pontchaillou,F-35000 Rennes, France*

Abstract. New use cases and the need for quality control and imaging data sharing in health studies require the capacity to align them to reference terminologies. We are interested in mapping the local terminology used at our center to describe imaging procedures to reference terminologies for imaging procedures (RadLex Playbook and LOINC/RSNA Radiology Playbook). We performed a manual mapping of the 200 most frequent imaging report titles at our center (i.e. 73.2% of all imaging exams). The mapping method was based only on information explicitly stated in the titles. The results showed 57.5% and 68.8% of exact mapping to the RadLex and LOINC/RSNA Radiology Playbooks, respectively. We identified the reasons for the mapping failure and analyzed the issues encountered.

Keywords. RadLex, LOINC, Mapping, Imaging, Clinical Data Warehouse

1. Introduction

The secondary use of imaging data, with AI methods, requires large scale data pooling and consequently interoperability in order to allow sharing data from different sources.

Healthcare data are defined using Interface Terminologies (ITs). In the biomedical field, an IT is commonly defined as "a systematic collection of healthcare related phrases (terms) that supports clinicians' entry of patient-related information into computer programs" [1]. However, the semantic interoperability in multi-center studies requires common, semantically defined terminologies. These Reference Terminologies (RTs) are defined as "terminologies designed to provide exact and complete representations of a given domain's knowledge, including its entities and ideas, and their interrelationships, and are typically optimized to support the storage, retrieval, and classification of clinical data" [2].

In a previous work, we proposed a pipeline to allow the integration, indexing and presentation of imaging data in our Clinical Data Warehouse (CDW) eHOP [3] via their metadata. These data come from the Picture Archiving and Communication System

[1] Corresponding Author, Pierre Lemordant, Univ Rennes 1, Inserm, LTSI UMR 1099, Rennes, France; E-mail: pierre.lemordant@univ-rennes1fr.

(PACS). Now, we want to map our imaging data to RTs to allow data sharing among different centers. We consider the RTs of RadLex and LOINC, whose coverage of ITs has been the subject of several studies [4-6].

In this work, we describe the details of mapping these local exam labels to RadLex and LOINC/RSNA terminologies and the barriers encountered.

2. Materials

As IT, we considered the labels describing imaging reports used at Rennes academic hospital over the last 18 years. For instance, the label "Scanner - Thorax sans IV" (Chest Computed tomography without intravenous injection of contrast agent) has been locally defined by the imaging department staff and is used in the radiology information system and in the electronic health records as a metadata of the imaging report for healthcare purposes. We queried these labels in our eHOP CDW which contains 1,467,000 imaging reports from more than 486,000 patients.

The Radiological Society of North America (RSNA) has created the RadLex Playbook [7], a standard system for naming radiological procedures that includes a list of 4,374 imaging procedure labels (terms) formed with elements of the RadLex ontology and identified by a RadLex Procedure ID (RPID). The RSNA and the Regenstrief Institute have been working together to create the LOINC/RSNA Radiology Playbook (L/R Playbook) [8] using a new information model to describe 6289 terms. This playbook uses LOINC ID as identifiers and provides correspondences between RadLex Playbook codes and LOINC codes. The RadLex Playbook has not been updated since 2018. The RSNA and Regenstrief Institute continue their collaboration to further develop the L/R playbook by adding new procedure codes identified only by a LOINC ID and not an RPID. In this work, we used version 2.71 of the L/R Playbook and version 2.5 of the RadLex Playbook.

3. Methods

First, we extracted the 200 most frequent imaging labels. This set represented our IT and covered 1,073,886 imaging exams (i.e 73.2% of all imaging exams in eHOP CDW). After removal of duplicates (e.g the same label but in upper case or with words separated by dashes instead of spaces), only 106 labels remained.

We then manually mapped our local labels to the RadLex and L/R Playbooks by considering all explicit information contained in the label. For example, although the expert (Y.G) knows that the "Ultrasound, intracranial vessels" exam is a Doppler exam, we did not map it to a RT term stating "Doppler" because this label did not specify the technique. Similarly, if the local term specified a reason for the exam, this reason must be mentioned in the RT term. Finally, the mapping was done with a single RT term, we did not combine several RT terms to describe a local term, unlike other approaches [5].

The mapping classification was rooted in previous classification proposals [9] and was as follows :

- **Exact match**: an RT code corresponded exactly to the procedure e.g. "Ultrasound - Abdomen-Kidney" perfectly matched "US ABD KIDNEY" (RPID1992);

- **Broader RT term issue**: the best RT candidate was broader in meaning than the local label. For instance, the L/R Playbook does not have a code containing all the elements from "CT - Chest Abdomen Pelvis Skull";
- **Narrower RT term issue**: some RT terms specify additional information that is not available in the local label, e.g. the RadLex Playbook always specifies information on the contrast agent in breast MRI and this did not allow finding an exact match for the local label "MRI - Breasts";
- **No exact match**: the local code used a concept that is not defined yet or never used in the RT. For instance, "Hemosiderosis" which is never used in the two playbooks (but is defined in the RadLex ontology (RID5203));

4. Results

4.1. Labels of the Interface Terminology

We observed much redundancy in our IT because the 200 most used labels represented 106 different imaging exams. This is explained by the fact that several teams of radiologists defined and modified this list of codes over the years. Among these labels, 80 specified only modalities and anatomic areas, 6 specified the contrast technique, 11 specified a procedure (e.g. "Densitometry"), 3 specified a reason (e.g. "Pulmonary embolism") and 2 specified a technique (e.g "Doppler").

By comparing these labels with the metadata that describe each acquisition made during the exam, we noted that the title was not always fully accurate. Indeed, sometimes, clinicians decide in the radiology room to make additional images than those scheduled, especially for X-rays exams in the context of trauma. For instance, procedures labeled as "X-ray - Wrist" contained a "wrist" acquisition but often also acquisitions targeting the "cervical spine", "elbow" or "clavicle".

4.2. Mapping of the Interface Terminology to Reference Terminologies

Table 1 shows the results of our manual mapping to the RadLex and L/R Playbooks.

Table 1. Outcomes of the manual mapping of the 106 local labels.

Mapping category	RadLex Playbook	LOINC/RSNA (L/R) Radiology Playbook
Exact match	61(57.5%)	73 (68.8 %)
Broader RT term issue	18 (17.0 %)	13 (12.3 %)
Narrower RT term issue	17 (16.0 %)	10 (9.4 %)
No exact match	10 (9.4 %)	10 (9.4 %)

The "Broader RT term issues" outcome occurred when the closest terms in the RT did not include all words to match the local label. We observed that the level of specification of RT terms can vary according to the modality, among other things. For example, the code "MR Lower Extremity Joint" exists in L/R Playbook, but there is no exact equivalent for "Ultrasound - Lower Extremity Joint".

In most cases, the "Narrower RT terms issues" outcome was due to the mention of the contrast agent in the RT labels, while this information was not specified in the local label. However, this information is not provided homogeneously in the RTs e.g. in the L/R Playbook, the terms "CT Chest", "CT Abdomen" and "CT Chest and Abdomen and Pelvis" (resp. 24627-2, 41806-1, 87869-4) do not mention contrast agent, whereas terms

describing CTs of "Chest and Abdomen" always specify the use of a contrast agent ("with", "without" or "without and with" resp. 42275-8, 42276-6, 42277-4). The expert review showed that in some cases, the imaging exam was done with and/or without a contrast agent in practice. However, as the local label did not explicitly specify this information, we could not perfectly match the label with an RT term, although the RT specifies the code that would allow the perfect match.

In the case of no exact match, the most common reason was the specification by the local code of the reason for the exam, or a procedure that was not mentioned in the two playbooks. For instance, "tuberculosis" or "cystography" are never used in the L/R Playbook (but exist in the RadLex ontology, RID34878 and RID29116). Another reason for the lack of match was the use of a new imaging technique that has not been added in the RT yet (e.g. the recently described EOS™ imaging system) [10]. Finally, in several cases, the French local label referred to an anatomic structure that was not referenced in the anatomical concepts of the RadLex ontology. For instance, the "Troncs supra aortiques" (supra-aortic trunk), which designates the brachio-cephalic artery, left common carotid artery and left subclavian artery has no exact equivalent in the RadLex ontology. Four of the ten "no exact match" cases identified (both playbooks) were explained by local labels referring to the "Troncs supra aortiques"

5. Discussion and Conclusions

The aim of this work was to identify the issues encountered in mapping an IT to a RT in the medical imaging field, and focused on the coverage of the local terminology by the two RTs by taking into account the whole content of local terms. This work tried to identify areas for improvement in IT and RT for data reuse and is, to our knowledge, the first work of this type using local French terminology.

As a limitation, our study was based on data from a single hospital and some of the identified issues may be specific to our institution. Moreover, we used a limited set of local terms and thus we might have missed problems linked to less common exams.

Other works about mapping of ITs to RTs in different domains mention the same problems of different granularity between ITs and RTs [4-6], or of terms that only have meaning within the local institution [6]. These observations show that ITs are primarily designed to be human-readable and highlight the importance of following good practice in creating ITs based on RTs [1,11]. This approach to creating ITs also permits the identification of terms that would be missing from the RT to efficiently evolve it [6].

Beitia *et al.* analyzed the CT procedure coverage by LOINC (before the unification of LOINC radiology procedures with RadLex) and RadLex terminologies [4] and obtained coverage rates of 70% and 75% respectively. The difference with our results can be explained by the stricter mapping method we used, the language barrier (e.g. "Supra aortic trunks"), the fact that we considered all modalities (by focusing only on CT, we obtained 75% and 66% of exact mapping to the L/R and RadLex playbooks respectively) and possibly by the construction method of our IT (e.g. the mention of "reasons for exam", such as "endometriosis" or "hemosiderosis", in local labels led to a "no exact match" with our mapping method).

Our results show that the mapping coverage was higher with the L/R Playbook than with the RadLex Playbook. However, we could note that the unification process is not completely achieved (e.g. the RadLex playbook term "XRAY BONE DENSITO" (RPID3335) does not yet have an equivalent in the L/R Playbook).

In this study, we could identify the elements to be taken into account concerning IT, RT, and data workflow in the healthcare system to develop a classification system of imaging exams that will allow optimal data integration and sharing among centers.

Acknowledgment

We would like to thank the French National Research Agency (ANR), for funding this work in the framework of the LabCom LITIS project (grant no. ANR-17-LCV1-0004).

References

[1] Rosenbloom ST, Miller RA, Johnson KB, Elkin PL, Brown SH. Interface terminologies: facilitating direct entry of clinical data into electronic health record systems. J Am Med Inform Assoc. 2006 May-Jun;13(3):277-88. doi: 10.1197/jamia.M1957. PMID: 16501181; PMCID: PMC1513664.

[2] Rosenbloom ST, Brown SH, Froehling D, Bauer BA, Wahner-Roedler DL, Gregg WM, Elkin PL. Using SNOMED CT to Represent Two Interface Terminologies. Journal of the American Medical Informatics Association. 2009 Jan;16(1):81-8.

[3] Madec J, Bouzillé G, Riou C, Van Hille P, Merour C, Artigny ML, Delamarre D, Raimbert V, Lemordant P, Cuggia M. eHOP Clinical Data Warehouse: From a Prototype to the Creation of an Inter-Regional Clinical Data Centers Network. Stud Health Technol Inform. 2019 Aug 21;264:1536-1537. doi: 10.3233/SHTI190522. PMID: 31438219.

[4] Beitia AO, Kuperman G, Delman BN, Shapiro JS. Assessing the performance of LOINC® and RadLex for coverage of CT scans across three sites in a health information exchange. AMIA Annu Symp Proc. 2013 Nov 16;2013:94-102. PMID: 24551324; PMCID: PMC3900124.

[5] Sandhu RS, Shin J, Wang KC, Shih G. Single-Center Experience Implementing the LOINC-RSNA Radiology Playbook for Adult Abdomen/Pelvis CT and MR Procedures Using a Semi-Automated Method. J Digit Imaging. 2018 Feb;31(1):124-132. doi: 10.1007/s10278-017-0016-0. PMID: 28842816; PMCID: PMC5788822.

[6] Peng P, Beitia AO, Vreeman DJ, Loo GT, Delman BN, Thum F, Lowry T, Shapiro JS. Mapping of HIE CT terms to LOINC®: analysis of content-dependent coverage and coverage improvement through new term creation. J Am Med Inform Assoc. 2019 Jan 1;26(1):19-27. doi: 10.1093/jamia/ocy135. PMID: 30445562; PMCID: PMC7587153.

[7] Radiological Society of North America, *Radlex playbook*, http://playbook.radlex.org

[8] The LOINC®/RSNA Radiology playbook. https://loinc.org/collaboration/rsna. Accessed January 2022.

[9] Humphreys BL, McCray AT, Cheh ML. Evaluating the coverage of controlled health data terminologies: report on the results of the NLM/AHCPR large scale vocabulary test. J Am Med Inform Assoc. 1997 Nov-Dec;4(6):484-500. doi: 10.1136/jamia.1997.0040484. PMID: 9391936; PMCID: PMC61267.

[10] Illés T, Somoskeöy S. The EOS™ imaging system and its uses in daily orthopaedic practice. Int Orthop. 2012 Jul;36(7):1325-31. doi: 10.1007/s00264-012-1512-y. PMID: 22371113; PMCID: PMC3385897.

[11] Griffon N, Savoye-Collet C, Massari P, Daniel C, Darmoni SJ. An interface terminology for medical imaging ordering purposes. AMIA Annu Symp Proc. 2012;2012:1237-43. PMID: 23304401; PMCID: PMC3540496.

Challenges of Trustable AI and Added-Value on Health
B. Séroussi et al. (Eds.)

doi:10.3233/SHTI220466

Deep SNOMED CT Enabled Large Clinical Database About COVID-19

Christophe GAUDET-BLAVIGNAC[a,b,1], Julien EHRSAM[a,b], Hugues TURBE[a,b], Daniel KESZTHELYI[a,b], Jamil ZAGHIR[a,b] and Christian LOVIS[a,b]
[a]*Division of Medical Information Sciences, University Hospitals of Geneva*
[b]*Department of Radiology and Medical Informatics, University of Geneva*

Abstract. In spring 2020, as the COVID-19 pandemic is in its first wave in Europe, the University hospitals of Geneva (HUG) is tasked to take care of all Covid inpatients of the Geneva canton. It is a crisis with very little tools to support decision-taking authorities, and very little is known about the Covid disease. The need to know more, and fast, highlighted numerous challenges in the whole data pipeline processes. This paper describes the decisions taken and processes developed to build a unified database to support several secondary usages of clinical data, including governance and research. HUG had to answer to 5 major waves of COVID-19 patients since the beginning of 2020. In this context, a database for COVID-19 related data has been created to support the governance of the hospital in their answer to this crisis. The principles about this database were a) a clearly defined cohort; b) a clearly defined dataset and c) a clearly defined semantics. This approach resulted in more than 28 000 variables encoded in SNOMED CT and 1 540 human readable labels. It covers more than 216 000 patients and 590 000 inpatient stays. This database is used daily since the beginning of the pandemic to feed the "Predict" dashboards of HUG and prediction reports as well as several research projects.

Keywords. COVID-19, Semantic Interoperability, SNOMED CT

1. Introduction

Since the beginning of 2020, the SARS-CoV2 pandemic is putting an important pressure on care systems throughout the world. In this context, the need for rapidly generating reliable evidence about a previously unknown disease is strong. As the volume and diversity of healthcare data is growing exponentially, the expectations on their use to reach this goal are high.

In Switzerland, five important epidemic waves hit the population since February 2020, the fifth being still ongoing in January 2022. [1] Those waves resulted in a dramatic increase of the demand for the healthcare system. In the canton of Geneva, the Geneva University Hospitals (HUG), a consortium of all public hospitals and multiple outpatient clinics, was designated to take care of COVID-19 inpatients. This resulted in an important shift in the activity of the hospital and required new tools to navigate and take decisions in these uncertain times. In the same period, an important rise in COVID-19 related research projects was observed, resulting in numerous demands for clinical data extraction and curation.

[1] Corresponding Author, Christophe Gaudet-Blavignac; E-mail: christophe.gaudet-blavignac@hcuge.ch

Large initiatives aimed toward healthcare data interoperability for research are already deployed in Switzerland, and international knowledge representation systems such as SNOMED CT are recommended to solve the semantic interoperability challenge of representing healthcare data. [2–4] However, this crisis highlighted several challenges limiting our ability to leverage this profusion of data to answer concrete questions related to public health measures or medical research. [5] Simple tasks such as counting cases, hospitalizations and deaths showed that if data exists, additional actions must be taken to make it technically and semantically actionable, such as the development of targeted ontologies. [6,7] Moreover, effective communication of this information is complex and must be adapted to efficiently disseminate evidence. [8]

The Division of Medical Information Sciences at HUG was mandated to create a database of COVID-19 related data. This database is used by the governance of HUG to support decision processes required by the pandemic. It is also used by several research projects and their respective IRB approvals. This work describes the creation of this database, and the approach taken to make clinical data related to COVID-19 actionable.

2. Methods

2.1. Design

The selected dataset was decided to be as inclusive as possible. Every patient tested for SARS-CoV2 regardless of the result or flagged in the electronic health record as COVID-19 positive or suspect of COVID-19 is included in the database. This approach is crucial to bring sustainability in a fast-evolving situation where knowledge and truth evolve every day. Indeed, the concept of positivity or suspicion of COVID-19 evolves with time, and each variant raises new challenges for detection and diagnostics. For example, the threshold for the positivity of a PCR is influenced by new detection methods such as salivary or nasal antigen testing and these add new categories of patients to be dealt with (positive to PCR but negative to antigen, etc.). Therefore, the most stable way of selecting cases to be included in the database is to include every patient tested for SARS-CoV2 as well as every suspicion or mention of COVID-19 in the Electronic Health Record (EHR).

For temporal coverage, it was decided that for each case included in the dataset, historical data available in HUG up to 2 years in the past would be extracted. This decision was based on the fact that, as new discoveries were made on the impact of SARSCoV2 infection, it became clear that patients were not equal in the severity of their disease. Therefore, as little was known about which preexisting condition, phenotypic or genotypic trait was relevant, it was decided that 2 years of full historical data would be extracted to be able to cluster cases and understand the underlying causes behind those variations.

The data model is derived from the source databases in HUG. It is structured around two key elements, the cases, which represent the patients, and the stays, whether inpatient or outpatient. The data is extracted from multiple sources depending on availability, the main one being the institutional data warehouse. When data is unavailable in the data warehouse, other sources are included, such as general consent information or structured radiologic reports. The database is updated hourly for a relevant subset of the data and every night for the complete dataset.

2.2. Semantic enrichment

Regarding the semantic interoperability of the database, a pragmatic approach was adopted. All data in the database would be encoded in an existing international standard, already encoded data would not be reencoded and SNOMED CT, when possible, would be preferred over other standards. This was because a clinical database is only actionable if the semantic of its content is clearly represented. The choice of SNOMED CT was made based on its coverage and the possibility to combine concepts.

To allow quick development of the database, only variables present in the dataset are encoded. Queries were created to generate the set of distinct clinical concepts present in the database. Those concepts are encoded by semantic experts using SNOMED CT. Each variable is linked to one or more SNOMED CT concepts according to the SNOMED CT Compositional Grammar. [9] For specific categories such as laboratory analysis or observations, French human readable labels are created.

Those semantic enrichments are integrated in the extract, transform and load process updating the database and updated regularly when new concepts appear in the data.

3. Results

3.1. Structure of the database

The development of the database began in March 2020. At the time of writing the database is still on and updating continuously. It now covers more than 200 000 patients and 590 000 inpatient stays. Its content is described in Table 1.

Table 1. Number of instances by category of data

Category	Number of instances
Patients	216 194
Laboratory results	35 069 657
Observations	161 785 476
Inpatient stays	590 610
Ambulatory consultations	479 973
Drug Prescriptions	3 278 824
Prescriptions other	26 178 856
Diagnosis codes	594 391
Procedure codes	566 476

The database contains 28 source tables containing raw data from each source. 29 materialized views contains the processed data, and the semantic encoding is stored in 10 mapping tables. Overall, the database's size is around 51 Gigabyte (Gb), the largest tables being observations (15Gb), laboratory analysis (8.5Gb) and medication orders (8.3Gb).

3.2. Variables and encoding

The encoding of the variables in SNOMED CT has been iteratively updated. It covers more than 28 000 variables and all the instances of laboratory analysis and observations.

Data already encoded in international standards such as ICD-10 [10] and ATC [11] have not been re-encoded due to the availability of mappings from those classifications to SNOMED CT. [12] Statistics about SNOMED CT encodings are listed in Table 2.

Table 2. Number of variables encoded in SNOMED CT

Variable	Number of SNOMED CT encodings
Laboratory analysis	4 826
Laboratory materials	1 089
Laboratory units	65
Observations	1 068
Formularies	125
Patient problems	13 680
Formularies' drop lists	4 223
Total	25 076

To simplify the navigation and understandability of its content, 1 540 human readable labels have been created for laboratory analysis and observations. Examples are shown in Table 3.

Table 3. Example of simplified labels for laboratory analysis and observations

Source data	French human readable label
SARS-CoV-2, ARN, PCR E-gene, ql (COBAS 6800)	Coronavirus - SARS-CoV-2 (COVID19), PCR
Ag-ADP 10µM	Agrégation plaquettaire
pv.upload.result.covid_19.positiveExternalDocumentIDs	Résultat test covid 19, externe positif
pv.ventilation.ohd.duration	Durée oxygénothérapie à haut débit

3.3. Database usage

Throughout the pandemic, this database has been used for multiple purposes. The main goal being to support the governance of the HUG in piloting the response to the pandemic. It is used daily to feed the "Predict" dashboards of HUG and to produce prediction reports.

Lastly, thanks to the inclusion of general consent information, the database, its simplified labels, and its semantic encoding have been used for data extraction for 5 research projects validated by the Institutional Review Board of HUG. Therefore, multiple types of users, from data scientists to caregivers doing research, can explore and understand the data.

4. Discussion

In this work, the creation of a database for COVID-19 related data in HUG is described. The three principles show important advantages for the creation of an actionable database for governance as well as for research.

By defining broad and simple selection criteria to define the cohort, covering 2 years of historical data, and encoding all data in international standards, this database can be used for multiple purposes. Its wide coverage both in term of included patients and selected variables helped to answer to the sustained flow of clinical questions raised by

the pandemic and to remain complete despite the variations in the definition of what a COVID-19 patient is.

Finally, the choice of SNOMED CT to encode every variable not already encoded in an international standard created a semantically interoperable representation of the information, usable to answer clinical questions. Moreover the ability of SNOMED CT to handle composition of codes through the Compositional Grammar was mandatory to correctly represent the 28 000 variables covered by the database. This semantic approach is compliant with the Swiss Personalized Health Network initiative that is currently deployed in Switzerland and is targeted at shaping the future of research.

5. Conclusions

Overall, this work showed that this three principles approach is pragmatic and can solve some of the challenges of health data interoperability. It succeeded in quickly filling the information needs of the hospital in a global crisis. A more systematic evaluation of this approach could further validate those results.

6. References

[1] COVID-19 Suisse | Coronavirus | Dashboard, (n.d.). https://www.covid19.admin.ch/fr/overview (accessed January 20, 2022).

[2] Gaudet-Blavignac C, Raisaro JL, Touré V, Österle S, Crameri K, Lovis C. A National, Semantic-Driven, Three-Pillar Strategy to Enable Health Data Secondary Usage Interoperability for Research Within the Swiss Personalized Health Network: Methodological Study. JMIR Med Inform. 2021 Jun 24;9(6):e27591. doi: 10.2196/27591. PMID: 34185008; PMCID: PMC8277320.

[3] SNOMED, SNOMED International, (2006). https://www.snomed.org/news-articles/snomed-ct-compositional-grammar-specification-and-guide (accessed November 6, 2017).

[4] Gaudet-Blavignac C, Foufi V, Bjelogrlic M, Lovis C. Use of the Systematized Nomenclature of Medicine Clinical Terms (SNOMED CT) for Processing Free Text in Health Care: Systematic Scoping Review. J Med Internet Res. 2021 Jan 26;23(1):e24594. doi: 10.2196/24594.

[5] Turbé H, Bjelogrlic M, Robert A, Gaudet-Blavignac C, Goldman JP, Lovis C. Adaptive Time-Dependent Priors and Bayesian Inference to Evaluate SARS-CoV-2 Public Health Measures Validated on 31 Countries. Front Public Health. 2021 Jan 21;8:583401. doi: 10.3389/fpubh.2020.583401. PMID: 33553088; PMCID: PMC7862946.

[6] Haber NA, Clarke-Deelder E, Feller A, Smith ER, Salomon JA, MacCormack-Gelles B, Stone EM, Bolster-Foucault C, Daw JR, Hatfield LA, Fry CE, Boyer CB, Ben-Michael E, Joyce CM, Linas BS, Schmid I, Au EH, Wieten SE, Jarrett B, Axfors C, Nguyen VT, Griffin BA, Bilinski A, Stuart EA. Problems with evidence assessment in COVID-19 health policy impact evaluation: a systematic review of study design and evidence strength. BMJ Open. 2022 Jan 11;12(1):e053820.

[7] de Lusignan S, Liyanage H, McGagh D, Jani BD, Bauwens J, Byford R, Evans D, Fahey T, Greenhalgh T, Jones N, Mair FS, Okusi C, Parimalanathan V, Pell JP, Sherlock J, Tamburis O, Tripathy M, Ferreira F, Williams J, Hobbs FDR. COVID-19 Surveillance in a Primary Care Sentinel Network: In-Pandemic Development of an Application Ontology. JMIR Public Health Surveill. 2020 Nov 17;6(4):e21434. doi: 10.2196/21434. PMID: 33112762; PMCID: PMC7674143.

[8] H. Turbé, V.G. Ruiz, M. Bjelogrlic, J. Rochat, and C. Lovis, Communicating on Multivariate and Geospatial Data supported by ergonomics criteria: COVID-19 case, in: 2020 Workshop on Visual Analytics in Healthcare (VAHC), 2020: pp. 4–16. doi:10.1109/VAHC53729.2020.00008.

[9] SNOMED CT Compositional Grammar Specification and Guide, (n.d.). https://www.snomed.org/news-articles/snomed-ct-compositional-grammar-specification-and-guide (accessed November 3, 2017).

[10] Classification of Diseases (ICD), (n.d.). https://www.who.int/standards/classifications/classification-of-diseases (accessed January 4, 2021).

[11] WHOCC - Home, (n.d.). https://www.whocc.no/ (accessed December 14, 2020).

[12] SNOMED CT maps, *SNOMED*. (n.d.). https://www.snomed.org/snomed-ct/Use-SNOMED-CT/maps (accessed January 20, 2022).

Challenges of Trustable AI and Added-Value on Health
B. Séroussi et al. (Eds.)

doi:10.3233/SHTI220467

Evaluating the Relevance of Virtual Drugs

Yvon K. AWUKLU[a,1], Vianney JOUHET[a,b], Sébastien COSSIN[b], Frantz THIESSARD[a,b], Romain GRIFFIER[a,b] and Fleur MOUGIN[b]
[a] *Bordeaux Univ. Hospital, Public Health Unit, Medical Information Department, F-33000 Bordeaux, France*
[b] *Univ. of Bordeaux, Inserm UMR 1219, Bordeaux Population Health Research Center, team ERIAS, F-33000 Bordeaux, France*

Abstract. Information about drugs is numerous and varied, and many drugs can share the same information. Grouping drugs that have common characteristics can be useful to avoid redundancy and facilitate interoperability. Our work focused on the evaluation of the relevance of classes allowing this type of grouping: the "Virtual Drug". Thus, in this paper, we describe the process of creating this class from the data of the French Public Drug Database, which is then evaluated against the codes of the Anatomical Therapeutic Chemical classification associated with the drugs. Our evaluation showed that 99.55% of the "Virtual Drug" classes have a good intra-class consistency.

Keywords. drug, virtual drug, semantic interoperability

1. Introduction

Information about drugs is numerous and varied, and many drugs like generic medicines share the same information. Therefore, a collection of drugs sharing common characteristics can be useful to avoid describing the same information several times, but also to contribute to the interoperability of the drug circuit and thus facilitate the international sharing of drug information [1].

Efforts are being made to aggregate information on drugs. Indeed, the SNOMED CT (Systematized NOmenclature of MEDicine Clinical Terms) and RxNorm, two major terminologies in the biomedical field, propose a representation of classes grouping drugs [2,3]. These are respectively the "Clinical Drug" and the "Semantic Clinical Drug", which propose a representation of drugs according to a combination of "active substance - strength in active substance - pharmaceutical form" (e.g. Product containing precisely paracetamol 500 milligram/1 each conventional release oral tablet [SCTID = 322236009], acetaminophen 500 MG Oral Tablet [RxCUI = 198440]) and bring together all drugs sharing these characteristics in a single concept. In France, the non-profit association Medicabase[2], which aims to build and make available semantic resources concerning medication, has developed a database of virtual drugs. These virtual drugs group similar drugs composed of the same active substance(s), the same strength in active substance and the same pharmaceutical form (e.g. Paracétamol 500 mg comprimé [id_MVN=MV00002306]).

[1] Corresponding Author, Yvon K. AWUKLU; E-mail: yvon.awuklu@outlook.com
[2] https://www.medicabase.fr/

Our objective was to assess the homogeneity of drug indications in each virtual drug class. For this purpose, we have created this class since the groupings of drugs as virtual drugs made by Medicabase have not yet been published. Thus, in this article, we will describe the process of creating and evaluating a class that aggregates drug information: the "Virtual Drug".

2. Materials and Methods

Grosjean et al. have defined the virtual drug as [4]: *"a drug that combines drug's brand names having: (i) the same active ingredient(s) or salts of the active ingredient(s) that are clinically equivalent in terms of iatrogenic risks; (ii) the same strengths of active ingredients in active base; and (iii) a galenic form considered clinically equivalent from the point of view of iatrogenic risks."*

The virtual drug "Paracetamol 500 mg, tablet" thus includes especially the products "DAFALGAN 500 mg, scored effervescent tablet", "DOLIPRANE 500 mg, tablet" and "EFFERALGAN 500 mg, orodispersible tablet" because they are equivalent according to this combination.

To build the "Virtual Drug" class, we used the French Public Drug Database (BDPM[3]) that provides online data on the composition of active ingredients, the strength of these active ingredients and the pharmaceutical form of the drugs marketed in France.

2.1. Implementing the Virtual Drug class

Before implementing this class, a preliminary step was necessary. It consisted in: (i) normalizing the pharmaceutical forms of the drugs, and (ii) normalizing and converting the strengths of the active ingredients. The normalization of pharmaceutical forms was achieved by the automatic grouping of forms sharing the same root (e.g. "scored film-coated tablet" and "scored effervescent tablet" have been grouped under the form "tablet"). With regard to the units of concentration of the active substances, an automatic standardization has been carried out in accordance with the Unified Code for Units of Measure terminology (UCUM[4]) (e.g. "µg" and "microgram" have been normalized to "ug"), as well as an automatic conversion of these strengths, if applicable (e.g. "1000 mg/2 mL" was converted to "500 mg/mL").

After this preliminary step, the "Virtual Drug" class was built, consisting of a group of drugs sharing the same composition in active ingredients, the same strength of these active ingredients and the same pharmaceutical form.

2.2. Evaluation

In an attempt to assess the overall consistency of this class, we hypothesized that any drug sharing the same composition in active ingredients, the same strength in these active ingredients and a similar pharmaceutical form should have the **same indications**.

[3] https://base-donnees-publique.medicaments.gouv.fr/

[4] A code system intended to include all units of measurement currently used in science, engineering and international trade (https://ucum.org/trac)

To test this hypothesis, we used the Anatomical Therapeutic Chemical classification system (ATC[5]), which gives the anatomical site of action, as well as the therapeutic class of a drug. Thus, ATC provides information concerning the indication of a drug from a generic perspective. The ATC codes were then automatically extracted from ROMEDI [5], Thesorimed[6] and the drug's summaries of product characteristics (SPCs[7]) (Figure 1). As the anatomical site and the therapeutic class of drugs can be identified in ATC codes of levels 2 to 5, ATC codes of the first level (e.g. N-Nervous system) were not included. The purpose of the evaluation was thus to determine whether or not the ATC codes associated with drugs belonging to a single "Virtual Drug" class were homogeneous (i.e. whether the drugs included in a virtual drug had all the same ATC code(s)).

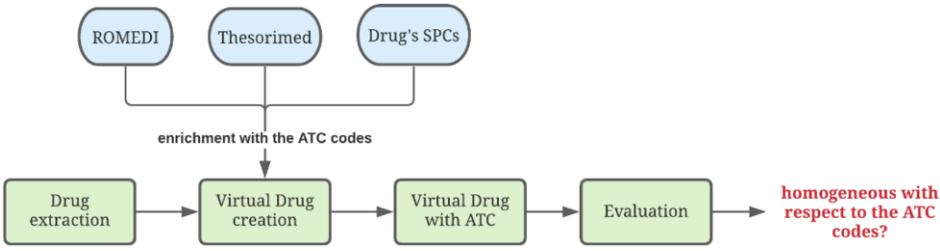


Figure 1. The "Virtual Drug" evaluation process

3. Results

The "Virtual Drug" class included 6,784 classes that were created from 15,318 drugs (Table 1). After pre-processing, 51 distinct drug forms and 1,367 distinct active substance strengths were obtained. The minimum number of drugs contained in a virtual drug is 1, while the maximum number is 41 (Table 2). "PARACETAMOL 500 mg : tablet" is the virtual drug that contained the most drugs (Table 3).

Table 1. Overview of the information about active substances, dosages and pharmaceutical strengths during the creation process of the "Virtual Drug" class

Drugs	Distinct Ingredients in drugs	Pharmaceutical forms		Strengths		Virtual drugs
		Before pre-processing	After pre-processing	Before pre-processing	After pre-processing	
15,318	3,531	414	51	2,181	1,367	6,784

Table 2. Information about the distribution of drugs in the "Virtual Drug" class

	Total	Min.	Q1	Q2	Mean	Q3	Max.
Virtual drugs	6,784	1	1	1	2.26	2	41
Virtual drugs with more than one drug	1,995	2	2	3	5.28	7	41

[5] https://www.whocc.no/atc/structure_and_principles/

[6] https://v3.prod-un.thesorimed.org/

[7] See for example, the SPC of "DOLIPRANE 500 mg, tablet" https://base-donnees-publique.medicaments.gouv.fr/affichageDoc.php?specid=63368332&typedoc=R

Table 3. Top 5 virtual drugs with the highest number of drugs

Virtual drug	Number of drugs
PARACETAMOL 500 mg: tablet	41
PANTOPRAZOLE 20 mg: tablet	39
DICLOFENAC SODIQUE 1 %: gel	36
OLANZAPINE 10 mg: tablet	34
IBUPROFENE 400 mg: tablet	33

The 1,995 virtual drugs comprising more than one drug included 10,529 (68.74%) drugs. Of these, 249 (12.48%) did not have an ATC code, including 240 (96.39%) classes containing only homeopathic drugs. Of the virtual drugs that had an ATC code, 1,555 (89.06%) contained drugs that all had at least one ATC code and 94.84% of their ATC codes were level 5 (being the most specific level).

The evaluation of these 1,555 "Virtual drug" classes showed that 111 (7.14%) of them are heterogeneous with respect to the ATC codes. The causes of heterogeneity were the following:

Source (46.0%). This case covers problems inherent in the source of ATC codes (i.e. incorrect ATC). For example, the virtual drug "ADAPALENE 0.1%: cream" has an ATC code D01AD03 that does not exist. The correct code is D10AD03.

Context (44.2%). This situation corresponds to the heterogeneity due to different ATC codes but having the same meaning in the "Virtual Drug" class in which they are found. For example, the virtual drug "PARACETAMOL | CODEINE 400 mg/20 mg: tablet" contains all three ATC codes (whose meaning in that specific class is the same):

- N02AA59: Codeine in combination with the exception of psycholeptics,
- N02AJ06: Codeine and paracetamol,
- N02BE51: Paracetamol in combination with the exception of psycholeptics.

Route (5.3%). This heterogeneity is related to the route of administration. For example, "ACETYLCYSTEINE 0.2 g / mL: solution" has two ATC codes R05CB01: ACETYL-CYSTEINE (respiratory route) and V03AB23: ACETYLCYSTEINE (venous route).

Granularity (3.5%). This cause includes problems due to the ATC code level provided for a drug. For example, "MIDODRINE HYDROCHLORIDE 2.5 mg: tablet" has the following two ATC codes: C01CA (level 4): Adrenergic and dopaminergic agents and C01CA17 (level 5): Midodrine.

Mode (0.9%). This heterogeneity is due to the mode of administration of a drug. For example, "ETHINYLESTRADIOL | DESOGESTREL 150 ug / 20 ug: tablet" has the ATC codes G03AB05: Desogestrel and ethinylestradiol (for sequential administration) and G03AA09: Desogestrel and ethinylestradiol (for fixed administration).

4. Discussion

As stated earlier, a class comparable to "Virtual Drug" is represented in SNOMED CT ("Clinical Drug") and RxNorm ("Semantic Clinical Drug") [2]. These two classes, although similar, do not incorporate pharmaceutical form grouping. In our study, this aggregation reduced the number of distinct pharmaceutical forms from 414 to 51, which resulting in better aggregation of information. Thus, it would be interesting to compare our virtual drugs to the RxNorm and SNOMED CT classes based on the other features we used to aggregate the information.

More than two thirds (70.59%) of the "Virtual Drug" classes contains only one drug. Despite this, the average number of drugs in the "Virtual Drug" and in his subgroup containing more than one drug is 2.26 and 5.28 respectively. These values reflect the aggregation capacity of this class. The five virtual drugs containing the largest number of drugs are dominated by analgesics, antisecretory drugs, nonsteroidal anti-inflammatory drugs and antipsychotics. These drugs represent the most prescribed drugs in France, which reflects the overall consistency of the "Virtual Drug" class.

To measure the internal consistency of each "Virtual Drug" subclass, it is necessary to assess whether we aggregated drugs that should not be aggregated and whether drugs that should be aggregated were not. The latter seems more difficult to measure because we do not yet have a method to do so. However, it was possible to assess the grouping of drugs that do not need to be aggregated. For this purpose, we used the ATC classification.

Virtual drugs containing only one drug logically have the maximum internal consistency. These classes were removed from our final evaluation dataset because they would have artificially overestimated the quality of the evaluation. The evaluation performed on the classes that could show low internal consistency identified only 7.14% of heterogeneous classes. Among the causes of heterogeneity, only those related to the route and the mode of administration (corresponding to only 0.45% of 1,555 virtual drugs used for the evaluation) are real sources of heterogeneity because they are directly linked to the drug. Therefore, the "Virtual Drug" class presents a good intra-class consistency regarding the indication.

5. Conclusions

In our study, we have shown that drugs belonging to the same virtual drug class have the same indications according to the ATC classification. As future work, we plan to verify this hypothesis with the specific indications of each drug, thus at a finer granularity level.

Acknowledgments

This work was partially supported by the INTENDED AI Chair (ANR-19-CHIA-0014).

References

[1] Rambaran KA, Huynh HA, Zhang Z, et al. The Gap in Electronic Drug Information Resources: A Systematic Review. Cureus. 2018;10:e2860.
[2] Nikiema JN, Bodenreider O. Comparing the representation of medicinal products in RxNorm and SNOMED CT – Consequences on interoperability. Proc 8th Int Conf Biomed Ontol ICBO 2019. 2019;6.
[3] Bodenreider O, Cornet R, Vreeman DJ. Recent Developments in Clinical Terminologies — SNOMED CT, LOINC, and RxNorm. Yearb Med Inform. 2018;27:129–139.
[4] Grosjean J, Letord C, Charlet J, et al. Un modèle sémantique d'identification du médicament en France. Proc Atelier IA Santé. 2019;9.
[5] Cossin S, Lebrun L, Lobre G, et al. Romedi: An Open Data Source About French Drugs on the Semantic Web. Stud Health Technol Inform. 2019;264:79–82.

Challenges of Trustable AI and Added-Value on Health
B. Séroussi et al. (Eds.)

doi:10.3233/SHTI220468

Preprocessing to Address Bias in Healthcare Data

Emel SEKER[a,b,1], John R. TALBURT[a], Melody L. GREER[b]
[a] *University of Arkansas at Little Rock, USA*
[b] *University of Arkansas for Medical Sciences, USA*

Abstract. Multimorbidity, having a diagnosis of two or more chronic conditions, increases as people age. It is a predictor used in clinical decision-making, but underdiagnosis in underserved populations produces bias in the data that support algorithms used in the healthcare processes. Artificial intelligence (AI) systems could produce inaccurate predictions if patients have multiple unknown conditions. Rural patients are more likely to be underserved and also more likely to have multiple chronic conditions. In this study, data collected during the course of care in a centrally located academic hospital, multimorbidity decreased with rurality. This decrease suggests a bias against rural patients for algorithms that rely on diagnosis information to calculate risk. To test preprocessing to address bias in healthcare data, we measured the amount of discrimination in favor of metropolitan patients in the classification of multimorbidity. We built a model using the biased data to test optimum classification performance. A new unbiased training data set and model were created and tested against unaltered validation data. The new model's classification performance on unaltered data did not diverge significantly from the performance of the initial optimal model trained on the biased data suggesting that bias can be removed with preprocessing.

Keywords. Data Bias, Artificial Intelligence, Underserved populations

1. Introduction

A rapid shift to data-centric processes punctuates recent history. Yet, healthcare has lagged behind other sectors like banking and retail, where machine learning and artificial intelligence were incorporated into the workflow decades ago. However, artificial intelligence is being incorporated into healthcare and healthcare-related settings such as insurance, population health, EHRs, disease screening, and clinical decision support systems (CDS) [1,2]. These computations are powered by data collected in the course of care. AI and real-world evidence can add value to patient care. Scrutiny will be required, however, as these tools are added into the healthcare process because our world is rife with examples of bias, and these are unfortunately captured in our data. We may be unknowingly propagating or even amplifying bias [3]. When an algorithmic prediction is based on biased information, it will result in biased predictions.

Comorbidity indices are common clinical data consumers used for risk adjustment based on patient characteristics [4]. Multimorbidity, having a diagnosis of two or more chronic diseases, is a risk factor for adverse clinical outcomes. Multi morbidity is known

[1] Corresponding Author, Emel SEKER, College of Medicine, Department of Biomedical Informatics, University of Arkansas for Medical Sciences, United States; E-mail: ESeker@uams.edu.

to have wide-ranging consequences and associations with poor outcomes, including decreased quality of life, psychological distress, more extended hospital stays, more postoperative complications, and a higher cost of care, ultimately resulting in higher mortality. Data from the 2018 National Health Interview Survey (NHIS) reported that 27.2% of US adults had multiple chronic conditions [5]. Clinical decision support systems need to incorporate complex information into risk analysis [2], but it isn't equally available for all populations. For example, when underserved Arkansas' rural communities were studied for diabetic neuropathy status, it was found that 79% had not been diagnosed with DPN (Diabetic Peripheral Neuropathy) among the patients with peripheral neuropathy symptoms [6]. Rural patients are, however, more likely to be underserved and also more likely to have multiple chronic conditions [7].

In this study, we examine bias in EHR data. We (1) measure the amount of discrimination, (2) de-bias the data by re-balancing class labels, and then (3) compare the pre-and post-processed modeling results.

2. Methods

An integrated data set was generated by appending zip code level data to 19,367 EHR records of patients with chronic diseases (asthma, diabetes, heart disease, congestive heart failure, coronary artery disease, heart attack, stroke) from the University of Arkansas for Medical Sciences Clinical Data Warehouse (AR-CDR) [8].

Patients are stratified by risk in order to receive care at the appropriate level of need and to produce the most optimal outcome for the patient. We modeled a simplified risk predictor to study how preprocessing data to remove bias impacts predictions. Because patients are evaluated for risk using primary vital signs (temperature, respiration, and heart rate), these were used as baseline clinical features along with demographic features (i.e., race, gender, geographic location).

The outcome variable was the class label *multimorbidity*. Urban, rural residence was designated as the sensitive attribute because multimorbidity decreased with rurality, as shown in Figure 1. This decrease suggests a bias against rural patients for algorithms that rely on diagnosis information to calculate risk. Geographic location data was appended in the form of Rural-Urban Commuting Area (RUCA) codes. Rural-Urban Commuting Area codes are indicators of the population level of a patient's geographic home location, which are generated using the United States Census Bureau data [9].

RUCA codes range from 1 to 9, indicating progressively more rural areas. Because the codes that are 6 or less indicate metropolitan areas and those equal to 7 or above indicate rural areas, they were binned into urban and rural categories accordingly.

The measure of bias against rural patients was measured as shown in Eq. (1) where: D is discrimination or bias, s is the sensitive attribute (rural residence condition), $\bar{s}$ is the sensitive (metropolitan residence condition). Resulting in D = 0.063, meaning that the data is biased in favor of metropolitan patients, 6.3% are assigned a more complicated multimorbid status than the rural patients.

$$D = \frac{\bar{s} \wedge multimorbidity}{\bar{s}} - \frac{s \wedge multimorbidity}{s} \tag{1}$$

Figure 1. Trendlines show a 6.3% bias in multimorbidity increase with age for urban but not rural patients.

The data was then separated into a 60%/40% training and validation set, and we built a model using the biased data to test optimum classification performance. To address the bias in the data, the class labels of multimorbidity and no-multimorbidity were changed for *M* ranked patient rows in the training set as shown in Eq. (2). Ranking was done using logistic regression to determine the probability of *multimorbidity* associated with each patient. Urban patients with the lowest probability of being classified as multimorbid were 'demoted' to *no-multimorbidity*, and rural patients with the highest probability of being classified as multimorbid were 'promoted' to *multimorbidity*. To maintain the balance between the two classes, promotion and demotion were done at the same time for each of 66 rows, as shown in Table 1. [10]

$$M = \frac{(s \; x \; \bar{s} \wedge multimorbidity) - (\bar{s} \; x \; s \wedge multimorbidity)}{s + \bar{s}} \tag{2}$$

A logistic regression model was then learned using the debiased training data and tested on the unaltered validation set. A logistic regression model was also learned using the unaltered training data with the sensitive attribute removed for comparison. We compared the results of these three methods in model classification performance.

Table 1. The data was debiased by rebalancing the multimorbidity class labels for M rows.

		Debiased Data - Number of Chronic Conditions	
Original Data - Number of Chronic Conditions		*no* Multimorbidity	Multimorbidity
	no Multimorbidity	5840	66
	Multimorbidity	66	13395

3. Results

We built a model using the biased data to test optimum classification performance. A new unbiased training data set and model were then created and tested against unaltered validation data. The new model's classification performance on unaltered data did not diverge significantly from the performance of the initial optimal model trained on the biased data suggesting that bias can be removed with preprocessing.

The initial bias or discrimination measurement was just over 6%, suggesting that patients from a metropolitan area had an advantage in that they would be evaluated at a higher risk stratification than their rural counterparts. This bias is unexpected because rural patients are more likely to have multiple chronic illnesses. Reclassifying ranked patients in a training set removed the discrimination and had a negligible impact on classification performance. Because removing sensitive attributes is also a potential solution, we compared these results with modeling using a training set with the sensitive attribute completely removed. This model was tested on unaltered validation data, and classification performance wasn't significantly different, as shown in Table 2.

Table 2. The AUC was measured and used to compare the classification performance of models built using data that was (1) unaltered, (2) debiased, and (3) had the sensitive attribute removed. The AUC was robust to debiasing techniques indicating that preprocessing can be applied without negatively impacting model performance in at least some cases.

	Unaltered	**Debiased**	**SA Removed**
Test	0.7033	0.7042	0.7031
Validation	0.7081	0.7088	0.7080

4. Discussion

Real-world clinical data is important for clinical decision-making and valuable when used within artificial intelligence algorithms. However, real-world data has everyday biases imprinted within it and can preserve and even amplify health disparities. Underserved rural populations are less likely to get needed healthcare due to distance, costs, and poor insurance coverage leading to underdiagnosis of illnesses even though they are more commonly affected by chronic conditions.

To study this bias problem within healthcare data, we have analyzed and preprocessed a real-world data set of patients with chronic conditions from geographically disparate locations. We have chosen to preprocess because it prepares the data set for any following modeling and does not need to be repeated for each new classifier. When correcting for bias it is essential to maintain the integrity of the data set for it to still be useful for prediction. We also tested the removal of the sensitive attribute

altogether. Each of these tests produces similar AUC results. Comparable AUC results indicate that classification bias can be removed while maintaining strong classification performance.

5. Conclusion

Our work has resulted in a desirable, in fact a necessary, outcome of stable prediction capability with reduced bias. Classification performance was not altered significantly in any of these cases, which suggests debiasing can be conducted without a drastic negative effect on predictive modeling. Although removing the sensitive attribute did not degrade classification performance, it is potentially detrimental because there are often multiple features associated with the sensitive attribute in question. This can continue to produce bias even in the absence of sensitive information.

Debiasing techniques may have unexpected downstream consequences that need to be evaluated. Further research is required in a broader range of institutions and sensitive attributes. We believe this is an essential first step in debiasing healthcare data used for algorithmic prediction and directly impacts patient health outcomes.

Funding Acknowledgment and Consent

Patients data used were obtained under IRB approval (IRB# 261145) at the UAMS.

References

[1] Makino M, Yoshimoto R, Ono M, Itoko T, Katsuki T, Koseki A, Kudo M, Haida K, Kuroda J, Yanagiya R, Saitoh E, Hoshinaga K, Yuzawa Y, Suzuki A. Artificial intelligence predicts the progression of diabetic kidney disease using big data machine learning. Sci Rep. 2019 Aug 14;9(1):11862. doi: 10.1038/s41598-019-48263-5. PMID: 31413285; PMCID: PMC6694113.

[2] Fraccaro P, Arguello Casteleiro M, Ainsworth J, Buchan I. Adoption of clinical decision support in multimorbidity: a systematic review. JMIR Med Inform. 2015 Jan 7;3(1):e4. doi: 10.2196/medinform.3503. PMID: 25785897; PMCID: PMC4318680.

[3] Seyyed-Kalantari L, Zhang H, McDermott MBA, Chen IY, Ghassemi M. Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in under-served patient populations. Nat Med. 2021 Dec;27(12):2176-2182. doi: 10.1038/s41591-021-01595-0. Epub 2021 Dec 10. PMID: 34893776; PMCID: PMC8674135.

[4] Alonso-Morán E, Nuño-Solinis R, Onder G, Tonnara G. Multimorbidity in risk stratification tools to predict negative outcomes in adult population. Eur J Intern Med. 2015 Apr;26(3):182-9. doi: 10.1016/j.ejim.2015.02.010. Epub 2015 Mar 6. PMID: 25753935.

[5] Boersma P, Black LI, Ward BW. Prevalence of Multiple Chronic Conditions Among US Adults, 2018. Prev Chronic Dis. 2020 Sep 17;17:E106. doi: 10.5888/pcd17.200130. PMID: 32945769; PMCID: PMC7553211.

[6] Wang W, Balamurugan A, Biddle J, Rollins KM. Diabetic neuropathy status and the concerns in underserved rural communities: challenges and opportunities for diabetes educators. Diabetes Educ. 2011 Jul-Aug;37(4):536-48. doi: 10.1177/0145721711410717. PMID: 21750334.

[7] CDC, Rural Health, in: *Preventing Chronic Diseases and Promoting Health in Rural Communities*, 2019.

[8] T.R. Institute, Clinical Data Repository (AR-CDR), in, 2022.

[9] USDA Economic Research Service, Rural-Urban Commuting Area Codes, in: *Economic Research Service*, U.S. Department of Agriculture, 2020.

[10] F. Kamiran and T. Calders, Classifying without discriminating, in: *2009 2nd international conference on computer, control and communication*, IEEE, 2009, pp. 1-6.

Challenges of Trustable AI and Added-Value on Health
B. Séroussi et al. (Eds.)

doi:10.3233/SHTI220469

Data Element Mapping in the Data Privacy Era

Romain GRIFFIER[a,b,1], Sébastien COSSIN[a,b], François KONSCHELLE[a,b], Fleur MOUGIN[b] and Vianney JOUHET[a,b]
[a] *Bordeaux University Hospital, Public health, 33000 Bordeaux, France*
[b] *Bordeaux University, Inserm U1219, Bordeaux Population Health, ERIAS team, 33000 Bordeaux, France*

Abstract. Secondary use of health data is made difficult in part because of large semantic heterogeneity. Many efforts are being made to align local terminologies with international standards. With increasing concerns about data privacy, we focused here on the use of machine learning methods to align biological data elements using aggregated features that could be shared as open data. A 3-step methodology (features engineering, blocking strategy and supervised learning) was proposed. The first results, although modest, are encouraging for the future development of this approach.

Keywords. data element, mapping, machine learning, LOINC

1. Introduction

Health data produced in the context of care can be reused for many purposes (phenotyping, research, etc.): this is the field of secondary use of health data. However, this reuse is complex (large volumes of data, compartmentalized data, etc.). One of the difficulties lies in the heterogeneity of the representation of medical concepts, called semantic heterogeneity. Different approaches can be used to reduce this heterogeneity. In particular, many efforts are focused on the alignment of local terminologies to international standards [1], such as the Logical Observation Identifiers Names and Codes [2] (LOINC®) used in many countries to encode biological data.

The Bordeaux University Hospital has implemented a clinical data warehouse (CDW) based on the i2b2 technology. The CDW integrates the data of patients who came at least once to the hospital since 2010, representing more than 2 million patients, 13 million hospital admissions and 2 billion observations. The CDW contains biology data integrated from two different biology software (TD-Synergy® until 2018 and Glims® since then) whose data are mostly centralized in the hospital's computerized patient record (DxCare®) resulting in a total of three biology sources. Each of the biology source software has its own local terminology and, within a biology source, several data elements (i.e. codes in the local terminology) may encode the same concepts, resulting in a high degree of semantic heterogeneity. One of the sources is partially aligned to LOINC. Thus, mapping these biological sources to each other would result in mapping local codes to LOINC.

[1] Corresponding author, Romain GRIFFIER ; E-mail: romain.griffier@chu-bordeaux.fr

With the development of machine learning methods, and in the context of strengthening personal data protection with the General Data Protection Regulation[2] (GDPR), many studies have raised privacy and security issues in AI methods [3].

Here, we propose to study the machine learning alignment of clinical datas through the example of biology records. Moreover, we aim at evaluating the performances of machine learning methods using aggregated features, thereby limiting the risk of compromising data privacy.

2. Methods

The proposed alignment methodology consists of three successive steps (Figure 1): 1) the data element selection and feature extraction, 2) the blocking strategy, and 3) a supervised classification.

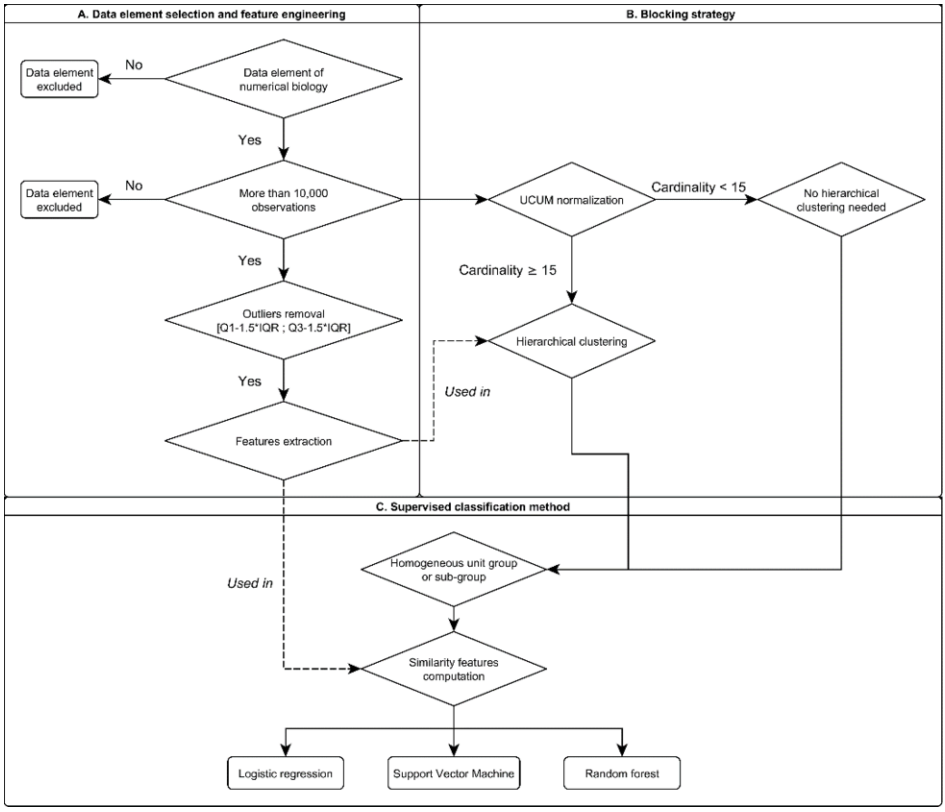


Figure 1. The 3-step alignment methodology

[2] https://gdpr.eu/

Data element selection and feature engineering. This first step consists in selecting the data elements[3] corresponding to the numerical biology with more than 10,000 observations available in the CDW (data-driven approach). For each selected data element, a data cleaning step with outlier removal was performed before computing the features. The following features were calculated: mean, standard deviation, median, quartiles, minimum, maximum, deciles, number of patients and stays, number of results by time of day (day or night), number of results above and below the norm. In addition, the sample distribution of each data element has been determined (using 1024 bins).

Blocking strategy. The final objective is to link similar identities (data elements) within the same data source and between several data sources. The number of possible comparisons is $\frac{n(n-1)}{2}$ where n is the sum of the cardinality of all data sources. To diminish the computational cost, it is necessary to have a blocking strategy [4] which limits the number of comparisons. The objective of this second step was therefore to constitute sub-groups of data elements in order to reduce the number of similarity features to be computed in step 3. This blocking strategy was based on:

1. The constitution of groups of homogeneous units on the basis of a standardization of units according to the Unified Code for Units of Measure[4] (UCUM) terminology.
2. Within the groups of data elements with a high cardinality (i.e. containing 15 or more data elements), an unsupervised clustering step using the hierarchical clustering (HC) method in order to form subgroups with lower cardinality.

Supervised classification. The third step was to compute the similarity features between all the data elements in each group resulting from the two previous steps. The following similarity features were calculated: difference in mean, difference in minimum, difference in maximum, difference in median, difference in quartile, difference in range and percent overlap of distributions. These similarity features were then used to train different supervised classification models: logistic regression, support vector machine (SVM) and random forest. The classification models were trained on a training sample (70%) and evaluated on a test sample (30%) using a gold standard of hand-crafted alignments by two experts in medical informatics (SC and RG).

Concerning the evaluation of the proposed method:

1. For the blocking strategy step, the evaluation only included gold standard concepts related to two or more data elements. Of these concepts, the percentage of those contained in a single homogeneous subgroup was assessed.
2. For the supervised classification method step, the evaluation was performed on the test set with recall, precision, F-measure, AUC and AUC_{PR}.

3. Results

Biological data integrated in the CDW represented 591,410,461 observations encoded in 170,933 data elements. The numerical biology corresponded to 475,117,464 observations (80.34%) encoded in 140,135 data elements (81.98%). After filtering out

[3] As defined by the ISO/IEC11179-3 standard. In our case, the data element corresponds to a code of a local terminology encoding a particular biological concept. Several data elements can encode the same concept (e.g. “HB001” and “HB002” are data elements that both correspond to the concept of Hemoglobin).

[4] https://ucum.org/ucum.html

Table 1. Results of the supervised classification models

	Threshold	Precision	Recall	F-measure	AUC_{PR}	AUC
Logistic regression	0.725	0.361	0.618	0.456	0.403	0.870
Support Vector Machine	0.710	0.391	0.476	0.430	0.393	0.800
Random forest	0.235	0.545	0.845	**0.663**	0.662	0.955

data elements with at least 10,000 observations available, 4,580 data elements (3.27%) remained representing 436,623,181 numerical observations (91.90%). Among these data elements, 1,421 (31.02%) were not associated with a unit (representing 89,465,534 observations). The others were associated with 153 different units.

After normalizing the units with UCUM, the data elements were grouped into 72 homogeneous unit groups (70 data elements could not be normalized with UCUM, representing 249,510 observations). The mean cardinality of these homogeneous unit groups was 44.29 data elements (sd=106.97). 42 homogeneous unit groups had cardinality less than or equal to 15 data elements and 30 homogeneous unit groups had cardinality greater than 15 data elements (mean=97.90; sd=151.32; maximum=709 for the percent unit).

The hierarchical clustering (HC) performed next on the 30 homogeneous unit groups with cardinality greater than or equal to 15 generated 277 clusters, with an average cardinality of 10.60 data elements (sd=30.94; maximum=336). Among the gold standard, considering biological concepts with at least 2 data elements, 95% of them were associated with data elements belonging to a single cluster.

Similarity features were computed in each of the 277 clusters (obtained by HC) and 42 groups of homogeneous unit groups with low cardinality, resulting in a total of 35,606 data element pairs. 3,756 (10.55%) of the data element pairs were associated with the same biological concept.

The results of the three supervised classification models are presented in Table 1. The best performing model was the random forest with an F-measure of 0.663, a recall of 0.845 and a precision of 0.545. Concerning the logistic regression and SVM models, the F-measures were respectively 0.456 and 0.430. The recall-precision curves for each model are presented in Figure 2 and found better overall performance for the random forest with AUC_{PR} of 0.662.

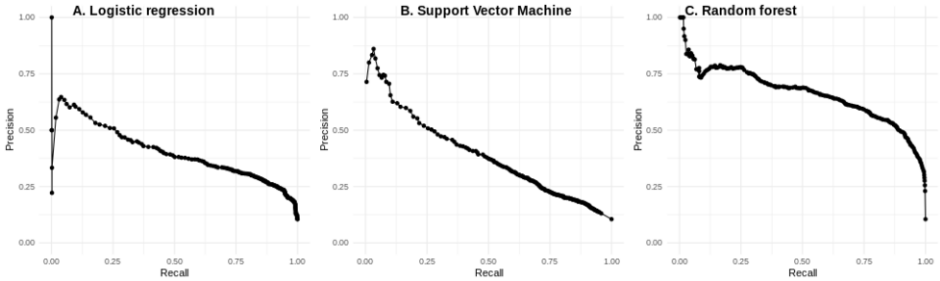


Figure 2. Recall-precision curves for the three classification methods

4. Discussion

We proposed a 3-step data-driven methodology to help achieve alignments between numerical biology data elements. The particularity of this work is the use of unsupervised

learning methods to implement a blocking strategy before training supervised classification models based on aggregated features that limit the privacy risk of re-identification based on trained algorithms.

Using HC to generate a blocking strategy reduced the cardinality of the homogeneous unit groups from 97.9 to 10.6 data elements without separating data elements of a same concept into different clusters. The performance of the HC was better than those obtained with the k-means method.

The supervised classification step yielded modest results. Using a random forest model gave the best results with an F-measure of 0.66 associated with a recall of 0.845 and a precision of 0.545 with a low threshold. The AUC was 0.955 in the context of a highly unbalanced data set (90/10). These results are slightly less good than those found in the literature [5], [6].

5. Conclusions

This preliminary work presents a 3-step method to align biological data elements using aggregated features that has obtained encouraging results. Further feature engineering work, including the addition of co-occurrence features, combined with semantic approaches, could optimize the performance of the proposed method, especially in the supervised learning step. An external validation step, using data from other healthcare institutions, will also be necessary to assess the generalizability of the method. Initiatives such as EHDEN[5] could provide a framework for implementing such an evaluation. Since this alignment method relies only on aggregated data at a very high level, sharing aggregate features related to LOINC concepts could help healthcare facilities to align their own local terminology with LOINC.

References

[1] Wade G, Rosenbloom ST. Experiences mapping a legacy interface terminology to SNOMED CT. BMC Med Inform Decis Mak. 2008 Oct;8 Suppl 1(Suppl 1):S3-3. doi: 10.1186/1472-6947-8-S1-S3.

[2] McDonald CJ, Huff SM, Suico JG, Hill G, Leavelle D, Aller R, et al. LOINC, a Universal Standard for Identifying Laboratory Observations: A 5-Year Update. Clin Chem. 2003 04;49(4):624-33. doi: 10.1373/49.4.624.

[3] Murdoch B. Privacy and artificial intelligence: challenges for protecting health information in a new era. BMC Med Ethics. 2021 Sep;22(1):122-2. doi: 10.1186/s12910-021-00687-3.

[4] Giang PH. A machine learning approach to create blocking criteria for record linkage. Health Care Manag Sci. 2015 Mar;18(1):93-105. doi: 10.1007/s10729-014-9276-0.

[5] Parr SK, Shotwell MS, Jeffery AD, Lasko TA, Matheny ME. Automated mapping of laboratory tests to LOINC codes using noisy labels in a national electronic health record system database. J Am Med Inform Assoc. 2018 Oct;25(10):1292-300. doi: 10.1093/jamia/ocy110.

[6] Nikiema JN, Griffier R, Jouhet V, Mougin F. Aligning an interface terminology to the Logical Observation Identifiers Names and Codes (LOINC®). JAMIA Open. 2021 Jun;4(2):ooab035-5. doi: 10.1093/jamiaopen/ooab035.

[5] https://www.ehden.eu/

Challenges of Trustable AI and Added-Value on Health
B. Séroussi et al. (Eds.)

doi:10.3233/SHTI220470

Temporal Medical Knowledge Representation Using Ontologies

Jacques HILBEY[a,b,1], Xavier AIMÉ[b] and Jean CHARLET[c,b]
[a] *Sorbonne Université, Paris, France*
[b] *Sorbonne Université, Sorbonne Paris Nord, INSERM, Laboratoire d'Informatique Médicale et d'Ingénierie des Connaissances en e-Santé - LIMICS, Paris, France*
[c] *Assistance Publique-Hôpitaux de Paris, Paris, France*

Abstract. Representing temporal information is a recurrent problem for biomedical ontologies. We propose a foundational ontology that combines the so-called three-dimensional and four-dimensional approaches in order to be able to track changes in an individual and to trace his or her medical history. This requires, on the one hand, associating with any representation of an individual the representation of his or her life course and, on the other hand, distinguishing the properties that characterize this individual from those that characterize his or her life course.

Keywords. Knowledge representation, biomedical ontologies, temporal representation, historical aspects

1. Introduction

Every living thing goes through all kinds of changes in its life course, and goes through episodes of health and disease. Medical knowledge reflects this, and medical textbooks define disease entities as sets of possible alternative courses [1].
These considerations, which are commonplace, lead to significant difficulties when one wishes to represent temporal medical information using ontologies based on Description Logics that admit only unary or binary relations, which restricts the possibilities of temporal indexing [2].

In the field of early detection and intervention in psychosis, which is our particular area of interest[2], temporal aspects are omnipresent, whether we consider the period of brain maturation in adolescence, the early phases of psychosis, the dynamic process of early psychosis, the duration of untreated psychosis, or schizophrenia as the last of a series of stages [3].

It is these problems of representing temporal aspects, recurrent in biomedical ontologies, that we address here by hypothesizing that the ontological commitment made in the foundational ontology is part of the solution. To this end, we present a foundational ontology and some examples of modeling of the situations considered.

[1] Corresponding Author, Jacques Hilbey; E-mail: jacques.hilbey@sorbonne-universite.fr.

[2] https://psy-care.fr/ - The clinical study to which the PsyCARE project will give rise will enable us to gather information allowing us to establish the knowledge graph of the patients in the study.

2. Methods

2.1. Petite Ontologie Fondationnelle (POF)

POF is a foundational ontology with 35 classes whose construction is based on two principles:

- a principle of individuation with respect to space and time, largely inspired by Zemach [4], which allows us to distinguish five kinds of individuals: objects, projects, events, properties, pieces of information;
- a mereotopological principle of division, assembly, situation of entities leading to distinguish: the whole, the part, the interface, the composition, the position.

The version we present here[3] is completed with a core ontology part simplified for our purposes.

2.2. Three-dimensionality and four-dimensionality

Representing in an ontology both objects that retain their identity through time (this chair, my nephew Louis, defined by the space they occupy - we speak of *three-dimensional* or 3D entities) and events (Louis' stay in Madrid[4], defined in space and time - we speak of *four-dimensional* or 4D entities) seems desirable: this is what natural language does, what "common sense" invites us to do, what important foundational ontologies in the biomedical domain do [5][6].

However, insisting on the properties that an object retains over time leads us to neglect, on the one hand, the fact that this object has not always existed and will not always exist, and, on the other hand, the changes that it will undergo during its existence. Furthermore, this representation raises a question of granularity: objects participate in events, but within these objects, smaller objects participate in smaller events. While this granularity of 3D entities is often considered and commonly represented as "Levels of Organization" or as a pyramid (e.g. Life's Complexity Pyramid), the equivalent does not exist for 4D entities, even though medical knowledge is always confronted with the articulation of different levels of event granularity.

From these two considerations, it follows that:

- a 3D entity must be considered from the point of view of the events in which it participates, but also of the events it hosts;
- a full-fledged representation of what we call an object should include, in addition to a class in the 3D branch of the ontology accounting for what persists in time, its counterparts in the 4D branch of the ontology to account for the fact that its existence is a limited process in time and space;
- the four-dimensional framework is more comprehensive than the three-dimensional framework and even if one wishes to retain 3D entities, it is more appropriate to focus on modeling 4D entities.

[3] https://framagit.org/jacqueshilbey/pof-mie-2022

[4] The example of Louis contracting an infection while in Madrid and subsequently receiving treatment is discussed below.

2.3. Implications on change representation

We can specify the notion of change from the six types of motion according to Aristotle [7]: generation, destruction, increase, decrease, alteration, change of place. Applied to a living organism, these types become: conception and death, quantitative changes, qualitative changes, and spatial displacement.

The representation of objects as continuants in time does not allow for the representation of their appearance and disappearance, whereas the representation of a life course includes the idea of an initial and a final space-time for it. On the other hand, a good reason for wanting to keep the three-dimensional representation is physical displacement: the notion of a change of place does not in itself carry the notion of a change in the entity that makes that spatial displacement. In the case of quantitative or qualitative changes, a more precise examination of the type of property involved in the change under consideration is necessary.

"Rigid" properties [8], which are only likely to be lost with the destruction of the entity, often appear in the ontology's taxonomy of classes (being a person). Descriptive properties such as height, weight, temperature, are no more likely to be lost by a material entity, but their value is likely to change over time.

The validity of the attribution of typical "anti-rigid" properties such as being in a *phase* (being an adolescent), holding a *role* (being a physician), or being in a *state* (being diseased), changes over time. They do not characterize identity over time, but rather a temporal part (or several temporal parts) of a life course. A phase can be naively defined as a temporal slice of a life course, although determining its boundaries can be a conundrum [9]. A role can be seen as a type of participation in certain events in the history of an entity. A state such as "being diseased" requires yet another model, insofar as the disease itself is an event, a process. This is why we propose to consider the course of the disease as an event whose spatiotemporal situation has as its spatial extension the continuant in time itself. This is justified by the fact that a continuant in time is precisely defined by the portion of space it occupies.

3. Results

We present a simple example: "Louis contracted an infection while in Madrid" and propose a representation (Figure 1) following the principles outlined above. An episode is defined by its spatiotemporal continuity and its narrative unity; a history is a composition of episodes. The representation of time *per se* is not deepened here: we just use two classes (*instant* and *temporal interval*) which are the temporal entities of the W3C Time Ontology in OWL[5].

SPARQL queries performed on the ontology are available in the same place as the ontology itself. They are accessible from the links in the footnotes.

Health Status[6] is a property whose value changes over time but whose assignment is always valid. The different instances that qualify the 3D entity are both instances of the "HealthStatus" class and instances of the subclasses of the corresponding value region. They are temporally indexed ("has beginning" and "has end" data properties) and can thus be queried according to their order or on a temporal interval.

[5] https://www.w3.org/TR/owl-time/

[6] https://framagit.org/jacqueshilbey/pof-mie-2022/-/blob/main/HealthStatus.rq

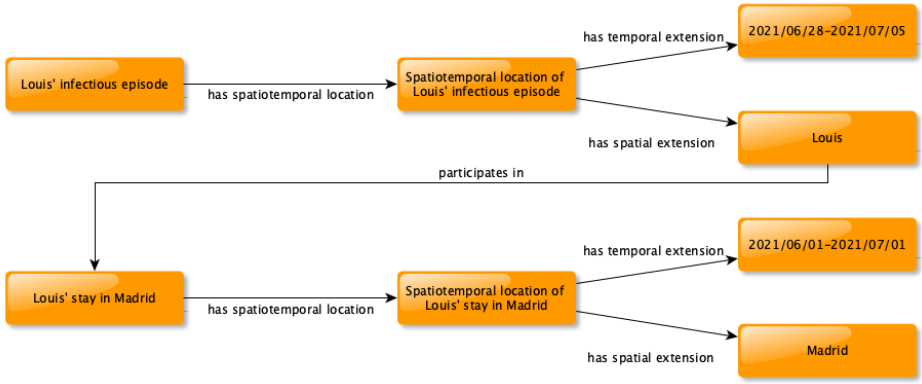


Figure 1. An object can participate in an event or be its spatial extension.

The **Medical History**[7]can be defined from the "internal" history[8] and the "external" history[9] of the individual as the whole made up of the events of medical interest of which it is the place and the events in which it takes part as a patient. This implies: on the internal history side, characterizing the events of medical interest either by creating a class or by assigning them a property for this purpose; on the external history side, creating as a sub-property of the property "participates in" a property "participates as a patient in", which carries the indication of the role.

As we have mentioned, it can be difficult to reach a consensus in the definition of a **Phase of Life**[10]. We can propose a minimal modeling, as the set of events of the internal or external history whose temporal extension has an intersection with a certain temporal interval. A more precise modeling would require that a life phase be defined at the level of the individual, from distinctive events.

4. Discussion and Conclusions

In this paper, we propose a new way to represent the temporal aspects of medical knowledge. We keep the representation of three-dimensional entities but limit it to what gives these entities an identity in time and their ability to participate as a whole in events. To the extent that they appear, change and disappear, these entities can also be considered as aggregates of processes.

The notion of history associated with a continuant in time already exists in a foundational ontology like BFO. Considering the granularity of 4D entities led us to subdivide it into external history, which is usually referred to as "life course", and internal history, which we will refer to, for lack of a better term, as "entire life process". In most medical specialties, the pathological processes are essentially episodes of this internal history. Taking it into consideration also makes sense in a pathophysiological approach to psychiatry.

[7] https://framagit.org/jacqueshilbey/pof-mie-2022/-/blob/main/InternalHistory.rq
[8] https://framagit.org/jacqueshilbey/pof-mie-2022/-/blob/main/InternalHistory.rq
[9] https://framagit.org/jacqueshilbey/pof-mie-2022/-/blob/main/ExternalHistory.rq
[10] https://framagit.org/jacqueshilbey/pof-mie-2022/-/blob/main/PhaseOfLife.rq

In the light of the above, we have been led to reconsider the attribution of properties to entities. Rigid properties can be represented as they are in BFO or DOLCE. But we substitute the consideration of anti-rigid properties of a 3D entity with that of temporal slices of internal or external history associated with that entity, whether we speak of phases or states; roles intervene as modalities of the entity's participation in events of its external history. Further research is needed to refine the modeling of stages, for example of a disease. Another research direction could be the establishment of SWRL rules in order to infer, for example, health status from an episode of disease.

We have mainly aimed here at the adequacy of the representation. As Burek et al. [2] remind us, there is no quality change representation choice that performs better than others in all dimensions. The aspects of scalability and extensibility will be investigated later, when integrating the data produced by the PsyCARE project.

Acknowledgements

This work has been supported by the French government's “Investissements d'Avenir” program, which is managed by the Agence Nationale de la Recherche (ANR), under the reference PsyCARE ANR-18-RHUS-0014.

References

[1] Hucklenbroich P. “Disease entity” as the key theoretical concept of medicine. J Med Philos. 2014 Dec;39(6):609-33.
[2] Burek P, Scherf N, Herre H. Ontology patterns for the representation of quality changes of cells in time. J Biomed Semantics. 2019 Oct 16;10(1):16.
[3] Krebs MO. Early detection and intervention: A change in paradigm. Ann Med Psychol (Paris). 2018 Jan;176(1):65-69.
[4] Zemach EM. Four Ontologies. Journal of Philosophy. 1970;67(8):231-247.
[5] Grenon P, Smith B. SNAP and SPAN: Towards Dynamic Spatial Ontology. Spatial Cognition and Computation. 2004 Mar;4(1):69-103.
[6] Gangemi A, Guarino N, Masolo C, Oltramari A, Schneider L. Sweetening ontologies with DOLCE. Berlin, Heidelberg: Springer; 2002. pp. 166–181.
[7] Aristote. Catégories. Paris : Éditions du Seuil ; 2002.
[8] A Formal Ontology of Properties. In: Dieng R, Corby O, editors. EKAW '00: Proceedings of the 12th European Workshop on Knowledge Acquisition, Modeling and Management; 2000 Oct 2; Berlin, Heidelberg: Springer-Verlag; c2000. p. 97-112.
[9] Sawyer SM, Azzopardi PS, Wickremarathne D, Patton GC. The age of adolescence. Lancet Child Adolesc Health. 2018 Mar;2(3):223-228.

Challenges of Trustable AI and Added-Value on Health
B. Séroussi et al. (Eds.)

doi:10.3233/SHTI220471

Applying Goal-Oriented Modelling for Machine Learning Based Rehabilitation Care

Carola GIßKE[a,1], Jin LIU[b] and Kai GAND[a]
[a] *Technische Universität Dresden, Faculty of Business and Economics, Research Group Digital Health, Germany*
[b] *Information Processing Engineering, FZI Forschungszentrum Informatik, Germany*

Abstract. Virtual coaches can support patients who need continuous rehabilitation due to an acute illness in the home environment. These coaching systems have to give medically correct instructions on the one hand and on the other hand respond individually to the patient. Hereby, machine learning algorithms could enable the adaptation and personalization of the rehabilitation process. In order to capture the necessary medical knowledge in a structured form and let the system technically make use of it, approaches of conceptual modelling have proved to be effective. On the basis of a virtual coaching scenario, we demonstrate how such a coaching application could be conceptually structured with the help of the goal-oriented modeling language i* in comparison to BPMN as process modelling approach and how machine learning algorithms could be implemented.

Keywords. Goal-oriented modeling, e-health, virtual coaching, machine learning

1. Introduction

A suitable rehabilitation after an acute episode or a chronic disease is important for the affected people in order to live independently and enhance their quality of life, especially in their home environment. Virtual coaches (VCs) could help these patients to engage in personalized rehabilitation programs [1]. In order to provide individualized rehabilitation, on the one hand coaching systems have to accompany the patient with medically correct instructions and on the other adapt flexibly to patient preferences. Process modelling is a well-known method to provide clinical knowledge in technical systems [2]. Machine learning approaches promise possibilities to make rehabilitation more flexible through continuous learning [3]. However, regarding the personalization of this clinical knowledge by means of machine learning, process models could limit the extent of adaptability due to their deterministic character. Therefore, more flexible approaches could be considered for these scenarios in order to foster the machine learning process.

In the following, a virtual coaching scenario for home rehabilitation will be used to analyze how a machine learning process based on goal-oriented knowledge modelling could be implemented. The following chapter describes the underlying use case and the steps from knowledge modelling to machine learning implementation. Chapter 3

[1] Corresponding Author, Carola GIßKE Technische Universität Dresden, Faculty of Business and Economics, Research Group Digital Health, 01062 Dresden, Germany; E-mail: carola.gisske@tu-dresden.de.

presents the results of the concrete approach. The last chapter concludes with a discussion and proposes further research steps.

2. Methods

The underlying use case is embedded in the EU-funded *vCare*[2] project which addresses the rehabilitation of chronic disabilities caused by an acute disease that leads to bodily limitations. A VC accompanies the patients during the rehabilitation at home by monitoring behavior and motivating the execution of practical exercises. Therefore, a dedicated technical architecture has been implemented in order to enable the VC solution. The foreseen procedures within require the modelling of medical knowledge necessary for the coaching process, first. This will be done as a clinical pathway [4] using an extended version of the standardized modelling language *BPMN*[3] and transformed into the machine-readable *HL7 FHIR* resource *PlanDefinition*. This initial and general pathway information will be forwarded to a professional portal, where a clinician can personalize pathway parameters for each patient and enrich the FHIR resource with individual values. From this, a wrapper for clinical pathways is used to generate an instance of the underlying ontology, which will then be processed in order to personalize the initial patients' pathway with machine learning algorithms using context information detected from system components like sensor data [5]. In particular, the reinforcement learning algorithm, contextual bandits, is applied to personalize rehabilitation for the patient. Through iterative interactions with the patient, the algorithm learns the mapping from patient states to appropriate recommendations [3].

Based on this approach, the knowledge modelling shall be replaced by a goal-oriented approach, namely the *i** language as a widespread framework [6]. Therefore, an exemplary rehabilitation scenario which describes the VC activity of monitoring the patients' daily activity and providing appropriate feedback will be used as single case study and thereby, transformed from BPMN (see Figure 1) to i* in order to see if the semantic modelling concepts of the different approaches can be matched. An attempt is then made to map the identified modelling concepts to the FHIR resource PlanDefinition. Finally, it will be examined whether this has an impact on the machine learning process.

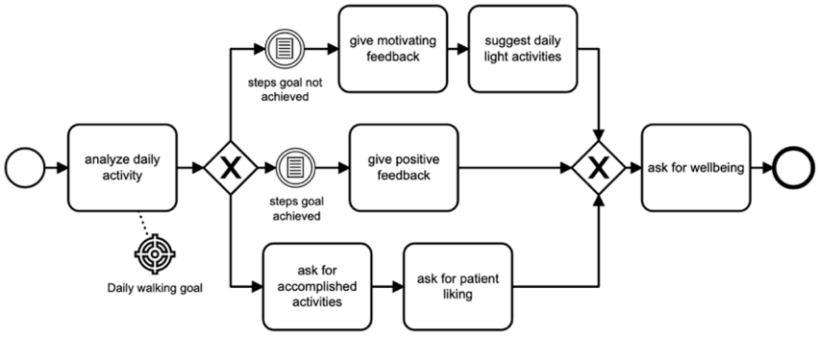


Figure 1. Exemplary virtual coaching use case represented with BPMN.

[2] See: https://vcare-project.eu

[3] See: https://www.bpmn.org

3. Results

The stated use case scenario and therefore, the underlying rehabilitation process information could be modelled using the i* language (see Figure 2). Compared to the initial process modelling with BPMN, the most noticeable difference is the modelling of different involved actors which in this use case includes, next to the VC, the patient's perspective. Additionally, the differentiation between *goals* and *softgoals* and their *contributions links* as well as the concept of *resources* that link actions between the actors in the system is new in this use case setting and needs to be considered.
The next step is to see whether the new concepts can be mapped to the FHIR resource PlanDefinition. For the system execution only the activities of the VC need to be taken into account, as the patient activities are monitored and thus, can be neglected during further processing. Likewise, the i* concept of resources does not need to be explicitly mapped, as this corresponds to the data generated by the system from involved components and with which the pathway information is enriched while processing.

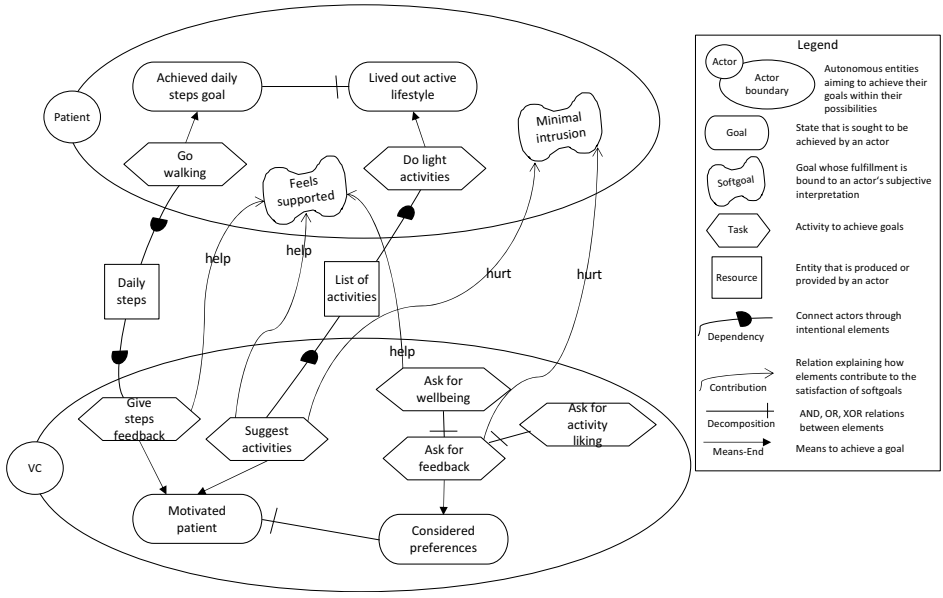


Figure 2. Virtual coaching use case represented with i*.

Table 1 compares the mapping of BPMN and i* concepts of the demonstrated use case to the FHIR resource PlanDefinition[4]. Differences are highlighted in cursive print.

Table 1: Different application of the FHIR resource *PlanDefinition* attributes to BPMN and i* concepts

FHIR attributes used with BPMN	**Description**	**FHIR attributes used with i***	**Description**
PlanDefinition.action.title	Actions represent the intended coaching activities performed by the VC, e.g. "ask for patient liking"	PlanDefinition.action.title	Actions represent the intended coaching activities performed by the VC, e.g. "suggest activity"

[4] See: https://www.hl7.org/fhir/plandefinition.html

PlanDefinition.action.code	A coding system is used to trigger system services, e.g. questionnaires or feedback	PlanDefinition.action.code	Could be adopted for the same reason as with BPMN
PlanDefinition.action.condition	Conditions describe criteria that will start the execution of an action, e.g. "steps goal achieved"	*PlanDefinition.action.selectionBehaviour*	Could be used to represent decomposition links, e.g. "AND" with FHIR attribute "all"
PlanDefinition.action.relatedAction.relationship	Used to describe the sequential action flow, e.g. before-start or after-end	*PlanDefinition.action.requiredBehaviour*	Could be used to represent contribution links, e.g. "hurt" with FHIR attribute "could"
PlanDefinition.goal.description	Used to define goals that should be achieved with or during actions, e.g. "daily walking goal"	PlanDefinition.goal.description	Could be adopted for the same reason as with BPMN, e.g. "achieved daily steps goal"
PlanDefinition.goal.target.measure	Defines the actual parameter value that needs to be tracked, e.g. number of steps	PlanDefinition.goal.target.measure	Could be adopted for the same reason as with BPMN
		PlanDefinition.goal.category	Could be used to differentiate between goals and softgoals
		PlanDefinition.action.goalID	Could be used to indicate which goals the action supports

The most significant difference between the attributes used is the representation of the relationship between the activities, which is in the nature of the two modelling languages. While the activities of BPMN follow an execution sequence, the activities in i* are flexibly selectable and their execution is based on the fulfilment of the defined goals. For the underlying architectural setting, this would mean that the ontology must be adapted accordingly in order to adequately represent the relationship between the activities. The differentiation between goal categories is new to the initial setting. Yet, this offers the opportunity to integrate and use this differentiation in the reward function of the machine learning algorithm and would allow for a more differentiated handling of goals regarding their importance. The execution of activities would therefore not be bound to conditions, but rather to how much the activity supports the set goals. Apart from that, this has no effect on the functioning of the machine learning algorithm. With goal-oriented modelling, the algorithm actually would have more possibilities and flexibilities to achieve rehabilitation goals. In the setting with BPMN, the learning process is bound to conditions when executing and suggesting activities and is dependent on related activities. With goal-oriented modelling, recommendations could be set depending on the observed states and iteratively it could be assessed which activities best support the rehabilitation goals set.

4. Study Limitations and Conclusions

Although goal-oriented modelling as a basis for the machine learning process could have the described advantages, its use also poses some challenges. First, the outcome of the

recommendations by the VC is not predictable as it is with BPMN. This in turn would require a lot of trust in the system by clinicians and patients as the system must be safe from a medical point of view. Also, differentiating the various goals in terms of the importance of achieving them requires medical expertise and therefore, close interaction with clinicians in the initial set-up of the system. The comparison of BPMN and i* in the given example suggests that i* is much more complex in representation than BPMN and thus, could be difficult as a basis for communication between technicians and medical professionals. Another aspect, which has not been considered yet, is the representation of time constraints. While the standard of BPMN provides a timing event in order to consider dates, repetitions or durations, this concept is not explicitly represented in the i* language. Since the architectural setting of the underlying rehabilitation scenario as the basis for the machine learning process does not change for both modelling languages considered, only the ontology and the transformation into the FHIR standard have to be slightly changed, a combined scenario could be considered. This could be designed in such a way that process modelling is first used as the structural basis for the system, on the one hand to give the clinical managers security in the ML process through deterministic and rule-based system behavior, and on the other hand to familiarize the patient with the system through recurring processes. Subsequently, the ML process could be implemented on the basis of goal-oriented modelling in order to adapt the process more individually to the patient and to personalize the coaching activities.

In the present work, it could be shown on the basis of a virtual coaching scenario for home rehabilitation that the goal-oriented modelling language i* can be used on the one hand to first map the coaching knowledge necessary for the system and then to transfer it to the HL7 FHIR standard in order to integrate the mapped information into an ML process. Further research should investigate the actual implementation of the proposed approach in order to analyze its potentials and acceptance compared to the process modelling approach. Additionally, the consideration of the timing perspective should be investigated to see how timing constraints could be considered in the goal-oriented modeling approach.

Acknowledgement

This project has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement No 769807.

References

[1] Tropea P, Schlieter H, et al. Rehabilitation, the Great Absentee of Virtual Coaching in Medical Care: Scoping Review. Journal of Medical Internet Research. 2019;21(10):e12805.
[2] Mincarone P, Leo CG, et al. Standardized languages and notations for graphical modelling of patient care processes: a systematic review. Int J Qual Health Care. 2018 Apr;30(3):169-177.
[3] Philipp P, Merkle N, Gand K, Gißke C. Continous support for rehabilitation using machine learnings. it – Information Technology. 2019;61(5-6): 273-84.
[4] Kinsman L, Rotter T, James E, Snow P, Willis J. What Is a Clinical Pathway? Development of a Definition to Inform the Debate. BMC Medicine. 2010;8:31-3.
[5] Gand K, Böcking L, Kreiner K, Schlieter H, Burwitz M, et al. Technically Representing Clinical Knowledge for Rehabilitation Care. Studies in Health Technology and Informatics. 2021;281:570-4.
[6] Franch X, López L, et al. The i* Framework for Goal-Oriented Modeling. In: Karagiannis D, et al, editors. Domain-Specific Conceptual Modeling. Cham: Springer International Publishing; 2016; p. 485-506.

Challenges of Trustable AI and Added-Value on Health
B. Séroussi et al. (Eds.)

doi:10.3233/SHTI220472

Pattern-Based Logical Definitions of Prenatal Disorders Grounded on Dispositions

Mirna EL GHOSH [a,1], Fethi GHAZOUANI [a], Elise AKAN [a], Jean CHARLET [a,c] and Ferdinand DHOMBRES [a,b]

[a] *INSERM, Sorbonne Université, Univ. Sorbonne Paris-Nord, LIMICS, Paris, France*
[b] *Médecine Sorbonne Université, GRC-26, Service de Médecine Foetale, AP-HP, Hôpital Armand Trousseau, Paris, France*
[c] *AP-HP/DRCI, Paris, France*

Abstract. Biomedical ontologies define concepts having biomedical significance and the semantic relations among them. Developing high-quality and reusable ontologies in the biomedical domain is a challenging task. Pattern-based ontology design is considered a promising approach to overcome the challenges. Ontology Design Patterns (ODPs) are reusable modeling solutions to facilitate ontology development. This study relies on ODPs to semantically enrich biomedical ontologies by assigning logical definitions to ontological entities. Specifically, pattern-based logical definitions grounded on dispositions are given to prenatal disorders. The proposed approach is performed under the supervision of fetal domain experts.

Keywords. biomedical ontologies, Ontology Design Patterns, logical definitions, prenatal disorders, BFO, dispositions

1. Introduction

High-quality and reusable ontologies are fundamentals for developing relevant semantic applications [1]. In the biomedical domain, building such ontologies is a challenging task, especially as the size and complexity of the ontology increases [2]. Ontology Design Patterns (ODPs) [3] address the quality and reusability concerns by providing different types of patterns supporting ontology design [1]. They are encouraging to capture common modeling situations, help facilitate ontology development and avoid common mistakes [2]. ODPs, which do not depend on any specific representation language, are categorized into different types such as *Presentation*, *Reasoning*, *Content*, and *Structural*. This work considers *Content* ODPs (CPs) being very beneficial kind of patterns for ontology design, because they provide solutions to domain-oriented problems [3]. They aim to solve modeling issues regarding ontology content, either in the general or a specific domain of the study [4]. We are interested in CPs in the specific domain which is the modeling of logical definitions.

[1]Corresponding Author: Mirna El Ghosh, INSERM, LIMICS, Sorbonne Universite, Univ. Sorbonne Paris-Nord, Paris, France; E-mail: mirna.el-ghosh@inserm.fr.

Logical definitions aim to associate terms within an ontology under development with terms in external species-neutral ontological resources such as BFO[2], PATO[3], and RO[4]. These definitions allow cross-species mapping using automated semantic reasoning and support quality control [5] and classifications (is-a/subclass relationships inferences). In biomedical ontologies, logical definitions are managed mainly in HPO[5] [6]. Encoding logical definitions in ontologies under development is a challenging task. Thus, approaches supporting pattern-based ontology design (e.g., eXtreme Design (XD) [7,8] and DOS-DPs [9]) and extensible ontology development (e.g., MIREOT [10]) are tackled. Such approaches bring solutions to create/reuse and apply logical definitions as ontology patterns.

This work is part of the SUOG (Smart Ultrasound in Obstetrics and Gynecology) project[6]. In SUOG, an ontology-based decision support system for complex ultrasound diagnosis in obstetrics and gynecology is intended [11]. The SUOG ontology, which is under development, distinguishes two main sub-ontologies: *prenatal disorders* (e.g., cerebral midline anomaly, complete molar pregnancy, etc.) and *prenatal phenotypes* (e.g. absent right superior caval vein, absence of discrete gestational sac, etc.) that suggest one or multiple disorders. In previous work [12], prenatal phenotypes entities are defined logically by reusing quality-oriented CPs from HPO[7] [6]. In this study, we focus on assigning logical definitions to prenatal disorders using disposition-oriented CPs.

The remainder of this paper is organized as follows. Section 2 introduces the methods. The results are presented in section 3. We discuss the study in section 4. Finally, section 5 concludes the paper.

2. Methods

Our research's main objective is to logically define prenatal disorders using CPs. To achieve the goal, we are based on eXtreme Design (XD) [7,8] to create the required ontology patterns. Besides, we envisaged that the targeted pattern(s) will be grounded on dispositions. The concept of `disposition` is addressed in some upper-level ontologies such as Basic Formal Ontology (BFO) [13,14] and the Unified Foundational Ontology (UFO) [15]. In BFO, a `disposition` (BFO:0000016) is defined as a *realizable dependent continuant inherent in some independent continuant* [16]. Two main characteristics of dispositions are exposed: 1) *dependency* - the existence of a disposition (e.g., fragility) requires a *bearer* (e.g., a material object such as glass) which is independent (i.e., to exist, a bearer is not required) and having some physical makeup features (e.g., molecular structure); 2) *realization* - a disposition is realized in a *process* (e.g., glass breaking) based on triggering conditions (e.g., the glass being forcefully pressed) [14,17]. Thereby, to exist, dispositions require some qualities inherent in the bearers (called *categorical basis* [18,19,20]) which are, or would be, causally relevant to the manifestation of a disposition [21]. Examples of categorical basis are the `structure` or `morphology` of the

[2]Basic Formal Ontology, http://www.obofoundry.org/ontology/bfo.html
[3]Phenotype And Trait Ontology, https://github.com/pato-ontology/pato/
[4]Relation Ontology, http://www.obofoundry.org/ontology/ro.html
[5]Human Phenotype Ontology, https://hpo.jax.org/app/
[6]https://www.suog.org/
[7]Human Phenotype Ontology, https://hpo.jax.org/app/

disposition's bearer. In the medical domain, *allergy* is an example of a disposition inherent in specific components of the immune system of an organism and *allergic reaction* represents the realization of allergy [22].

In the SUOG ontology, prenatal disorders are prescribed as "malformations" or "anomalies" affecting fetal organs. The ontological analysis of disorders is debatable in the biomedical ontology engineering community. In this study we follow Scheuermann [23] and Rohl [19] approach that considers *disorders as physical basis of dispositions*. Based on these perspectives, the ontological requirements in SUOG and a list of Competency Questions (CQs) are defined. Examples of CQs are: (CQ1) what dispositions having prenatal disorders as material basis? (CQ2) What anatomical structure is the bearer of such dispositions? (CQ3) What qualities are the categorical basis of such dispositions? (CQ4) What is the bearer of the categorical basis? (CQ5) The disposition is realized in what process? (CQ6) Exist any participants in the process?

In the following, we give a formalization of the specifications of dispositions applied in the context of prenatal disorders using some conventional logical operators of first-order logic such as $\rightarrow$, $\exists$, and $\wedge$. In formalization, unary predicates are represented in bold, and relations are in italic and x, y, and z are variables. First, the specification of prenatal disorders as material basis of dispositions is formalized.

$$\textbf{prenatal disorder}(\mathrm{x}) \rightarrow \exists y\ (\textbf{disposition}(\mathrm{y}) \wedge \textit{is material basis for}(\mathrm{x},\mathrm{y}))$$

The specification of dispositions' bearers, representing the affected anatomical structures associated to disorders, is formalized as follows.

$$\textbf{disposition}(\mathrm{x}) \rightarrow \exists y\ (\textbf{anatomical structure}(\mathrm{y}) \wedge \textit{inheres in}(\mathrm{x},\mathrm{y}))$$

The specification of dispositions' categorical basis is formalized as follows.

$$\textbf{disposition}(\mathrm{x}) \rightarrow \exists \mathrm{y}\ (\textbf{quality}(\mathrm{y}) \wedge \textit{has categorical basis in}(\mathrm{x},\mathrm{y}))$$

In the SUOG ontology, a prenatal disorder is *suggested* by one or more prenatal phenotypes (defined as *qualities* captured by ultrasound mechanisms). Thus, following BFO's definition of *qualities as continuants* and analysis of *processes having continuants as participants*, we consider that phenotypes participate in the realization of dispositions (i.e., processes). The specification of dispositions realization is formalized as follows.

$$\textbf{disposition}(\mathrm{x}) \rightarrow \exists \mathrm{y}\ (\textbf{process}(\mathrm{y}) \wedge \textit{has realization}(\mathrm{x},\mathrm{y})) \wedge \exists \mathrm{z}\ (\textbf{prenatal finding}(\mathrm{z}) \wedge \textit{has participant}(\mathrm{y},\mathrm{z}))$$

These formalizations are considered to create `P`, a disposition-oriented pattern composed of six main clauses. `P` includes basic categories (concepts and relations) which are reused from existent validated ontologies and variables (*var*), which will be filled during pattern application. Examples of basic concepts are `disposition`, `process`, and `quality` reused from BFO. Examples of relations are *is material basis for* and *has categorical basis in* which are specializations (in the SUOG ontology) of *is basis for realizable* (RO_0004018) and *realizable has basis in* (RO_0004017) respectively.

```
P:'is material basis for' some ('bfo:disposition'
              and 'has realization' some ('bfo:process'
                        and 'has participant' some var)
              and 'inheres in' some var
              and 'has categorical basis in' some (var
                        and 'inheres in' some var))
```

3. Results

The disposition-oriented pattern is applicable to logically define the different categories, sub-categories, and specific classes of prenatal disorders. Figure 1 depicts examples of logical definitions assigned to `cerebral midline anomaly` and `complete molar pregnancy`. The application of the proposed pattern in the SUOG ontology is in progress. A total of 11 general categories and 45 sub-categories of prenatal disorders are logically defined. The generated logical definitions are verified against the CQs, evaluated, and validated by the fetal domain experts.

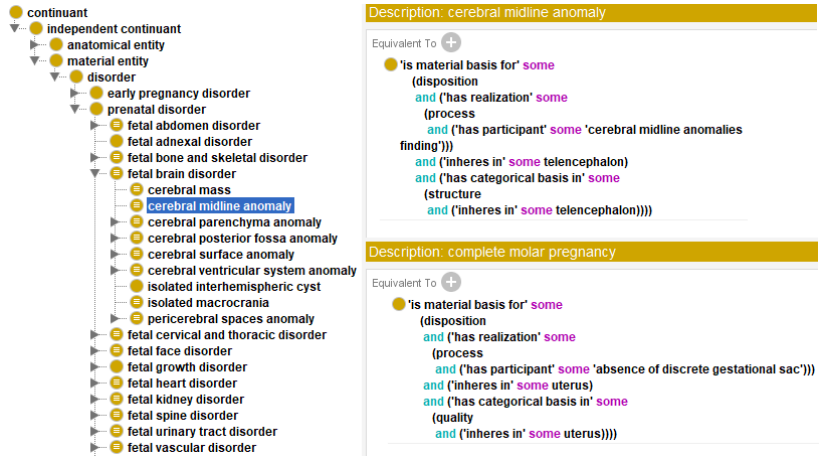


Figure 1. Examples of logical definitions assigned to prenatal disorders

4. Discussion

In this work, we have proposed a pattern-based approach for assigning logical definitions to prenatal disorders grounded on dispositions. By applying this approach, we aligned the SUOG ontology to upper-level ontological resources and enriched the ontology model semantically. Besides, using Content ODPs has simplified the ontology engineering process and permitted the extensibility of the ontology. In this study, the SUOG ontology is aligned to BFO [13]. In further work, we will analyze the definition of dispositions provided in UFO [15] and its applicability to build/enhance our pattern(s). On the conceptual level, in the proposed (general) pattern, we considered a single categorical basis for a disposition. Meanwhile, a disposition may have multiple categorical basis [17] which will be envisaged for defining the most specific prenatal disorders entities. Finally, concerning the participation of prenatal phenotypes in the realization of dispositions, the proposed pattern will be specialized to cover the cases where multiple prenatal findings are participating. For the best application of the disposition-oriented pattern(s), it is required that the fetal domain experts perform or supervise the fulfillment of variables for each prenatal disorder entity (e.g., 4 variables/disorder in `P`). The associated values, which are provided as *csv* file, represent the input data required for the automatic application of the patterns. This process risks delaying the pattern application to define the entire prenatal disorders entities (total 1132 in the current version of the SUOG ontology).

5. Conclusion

Building high-quality and reusable ontologies in the biomedical domain is challenging due to ontologies' increasing size and complexity. Ontology Design Patterns and ontology reuse approaches are tackled to overcome the challenges and simplify the ontology building process. This study discussed pattern-based logical definitions of prenatal disorders entities in the SUOG ontology. A proposed content pattern is grounded on BFO:disposition leading to promising results in modeling logical definitions.

References

[1] Blomqvist E, Sandkuhl K, Scharffe F, Svatek V, editors. Proceedings of WOP-2009, collocated with ISWC-2009. Washington D.C., USA; 2009.

[2] Mortensen JM, Horridge M, Musen MA, Noy NF. Applications of ontology design patterns in biomedical ontologies. In: AMIA; 2012. p. 643–652.

[3] Gangemi A, Presutti V. Ontology design patterns. In: Staab S, Studer R, editors. Handbook on Ontologies. Berlin, Heidelberg: Springer; 2009. p. 221-43.

[4] Hammar K. Content Ontology Design Patterns: Qualities, Methods, and Tools. Linkoping University; 2017.

[5] Köhler S, Bauer S, Mungall CJ, Carletti G, Smith CL, Schofield P, et al. Improving ontologies by automatic reasoning and evaluation of logical definitions. BMC Bioinformatics. 2011;12.

[6] Köhler S ea. Expansion of the human phenotype ontology (hpo) knowledge base and resources. Nucleic Acids Research. 2019;47.

[7] Presutti V, Daga E, Gangemi A, Blomqvist E. eXtreme Design with Content Ontology Design Patterns. In: Proceedings of WOP-2009, collocated with ISWC-2009; 2009. p. 83–97.

[8] Blomqvist E, Hammar K, Presutti V. Engineering Ontologies with Patterns: The eXtreme Design Methodology. In: Ontology Engineering with Ontology Design Patterns. IOS Press; 2016. p. 23-50.

[9] Osumi-Sutherland D, Courtot M, Balhoff JP, Mungall C. Dead simple OWL design patterns. J Biomed Semantics. 2017;8.

[10] Courtot M, Gibson F, Lister AL, Malone J, Schobe D, Brinkman R, et al. MIREOT: the Minimum Information to Reference an External Ontology Term. Nature Precedings. 2011.

[11] Dhombres F, Maurice P, Guilbaud L, Franchinard L, Dias B, Charlet J, et al. A novel intelligent scan assistant system for early pregnancy diagnosis by ultrasound: Clinical decision support system evaluation study. JOURNAL OF MEDICAL INTERNET RESEARCH. 2019;21.

[12] Ghosh ME, Ghazouani F, Birene B, Akan E, Charlet J, Dhombres F. Modeling Logical Definitions in Biomedical Ontologies by Reusing Ontology Design Patterns. In: ICBO; 2021. p. 56-62.

[13] Smith B, Kumar A, Bittner T. Basic Formal Ontology for Bioinformatics; 2005.

[14] Arp R, Smith B, Spear A. Building Ontologies With Basic Formal Ontology. Cambridge, MA: MIT Press; 2015.

[15] Guizzardi G, Benevides AB, Fonseca CM, Porello D, Almeida JPA, Sales TP. UFO: Unified Foundational Ontology. Applied Ontology. 2021.

[16] Arp R, Smith B. Function, Role, and Disposition in Basic Formal Ontology. Nat Prec. 2008.

[17] Toyoshima F, Barton A, Jansen L, Ethier JF. Towards a Unified Dispositional Framework for Realizable Entities. In: Proceedings of FOIS2021. IOS Press; 2021. .

[18] Goldfain A, Smith B, Cowell LG. Dispositions and the Infectious Disease Ontology. In: Formal Ontology in Information Systems, FOIS 2010; 2010. p. 400-13.

[19] Rohl J, Jansen L. Representing Dispositions. Journal of Biomedical Semantics. 2011;2(Suppl 4).

[20] Barton A, Grenier O, Jansen L, Ethier JF. The identity of dispositions. In: Proceedings of FOIS2018. IOS Press; 2018. p. 113-25.

[21] McKitrick J. Dispositional pluralism. Oxford University Press; 2018.

[22] Schulz S, Spackman K, James A, Cocos C, Boeker M. Scalable representations of diseases in biomedical ontologies. In: Bio-Ontologies 2010: Semantic Applications in Life Sciences; 2011. .

[23] Scheuermann RH, Ceusters W, Smith B. Toward an ontological treatment of disease and diagnosis. In: AMIA Summit Transl. Bioinforma; 2009. p. 116–120.

Challenges of Trustable AI and Added-Value on Health
B. Séroussi et al. (Eds.)

doi:10.3233/SHTI220473

Using Big Data to Identify Impact of Asthma on Mortality in Patients with COVID-19

Jinyan LYU[a,1], Wanting CUI[a] and Joseph FINKELSTEIN[a]
[a] *Icahn School of Medicine at Mount Sinai, USA*

Abstract. The goal of this paper was to assess if mortality in COVID-19 positive patients is affected by a history of asthma in anamnesis. A total of 48,640 COVID-19 positive patients were included in our analysis. A propensity score matching was carried out to match each asthma patient with two patients without history of chronic respiratory diseases in one stratum. Matching was based on age, comorbidity score, and gender. Conditional logistics regression was used to compute within each strata. There were 5,557 strata in this model. We included asthma, ethnicity, race, and BMI as risk factors. The results showed that the presence of asthma in anamnesis is a statistically significant protective factor from mortality in COVID-19 positive patients.

Keywords. COVID-19, Conditional Logistic Regression, Mortality

1. Introduction

The COVID-19 pandemic brought about 64,062,060 confirmed cases and resulted in 846,463 deaths in the United States [1]. Such a high number of positive patients significantly strained medical care delivery. One way to optimize healthcare resources is to identify potential severity of patients in advance. Patients with chronic respiratory diseases (CRD) who contracted SARS-CoV-2 may potentially suffer more severe consequences from COVID-19. A substantial proportion of CRD patients have asthma. A study from Scotland showed that adults with asthma had higher hospitalization, ICU admission, and death than those without asthma [2]. Our study was aimed at assessing whether mortality in COVID-19 positive patients is affected by a history of asthma in anamnesis. In this study, big data repository from a nation-wide collaborative was used to compare mortality rates in COVID-19 positive patients with history of asthma and COVID-19 positive patients without CRD.

2. Methodology

This study used the data from National COVID Cohort Collaborative (N3C) Data Enclave. N3C is a national wide COVID-19 clinical data hub that contains demographic and clinical characteristics of patients who have been tested for or diagnosed with

[1] Corresponding Author, Jinyan Lyu; E-mail: jinyan.lyu@mssm.edu.

COVID-19. The dataset contains over 10 billion rows of data, 3.3 million COVID test-positive patients, and 9.4 million total patients [3]. Due to the requirements of HIPAA, all analyses were performed on N3C's official data enclave website.

We included adult patients (age >= 18) who underwent COVID-19 testing between January 1, 2020 and April 30, 2021 and had valid test results. A COVID-19 test extraction was based on the validated LOINC codes. A COVID-19 test result was defined as positive if the result was reported as 'Positive' or 'Detected.' A test result was negative if the result was reported as 'Not detected' or 'Negative.' A COVID-19 negative patient was defined if all COVID-19 tests were negative. A COVID-19 positive patient was defined as having at least one positive test result. Death of patients deceased within 90 days after positive COVID-19 testing was attributed to COVID-19. The mortality rate in these patients was defined the patients' COVID-19 mortality rate.

For each COVID-19 positive patient, analyses included gender, age, race, ethnicity, BMI, and state of residence in US. Asthma was identified by ICD-10 code J45. We included asthma patients without other CRD. Patients without CRD were defined as having no documented COPD (J42-J44), asthma (J45), and lung cancer (C34) in the condition occurrence table. The age-adjusted Charlson comorbidity index score was calculated from patients' history using ICD-10 codes [4]. We divided age into four age groups 18- 35, 36-50, 51-75, and 76+ as age confounder. The Charlson comorbidity score was stratified into four levels (0-2, 3-4, 5-6, and 7 or higher) as a total comorbidity status. We removed patients with missing values or extreme outliers in age, gender, ethnicity, race, state, and BMI. All data was extracted from the condition occurrence table and then merged with independently identified patients in the patient table. PySpark was used to perform analyses of billions of medical records in N3C.

The analyses described in this publication were conducted with data or tools accessed through the NCATS N3C Data Enclave https://covid.cd2h.org and N3C Attribution & Publication Policy v 1.2-2020-08-25b supported by NCATS U24 TR002306. This research was possible because of the patients whose information is included within the data and the organizations (https://ncats.nih.gov/n3c/resources/data-contribution/data-transfer-agreement-signatories) and scientists who have contributed to the on-going development of this community resource [3].

2.1 Statistics model

We applied a random forester model to calculate feature importance. The result showed that age, total comorbidity score, and gender were the most important features 0.498, 0.300, 0.033, respectively. This result is in accordance with a recent paper analyzing CRD using a cluster analysis that showed that age, total comorbidity score, and gender were important factors [5]. We employed propensity-score matching method to reduce the effect of confounders. Age, total comorbidity score, and gender were used for matching. Using R package ‘matchit’, the nearest neighbor matching method matched two groups with a ratio of 1:2 (case: control) on age status, comorbidity status, and gender. All patients who only have asthma were assigned into the case group and matched accordingly with non-CRD patients who belonged to the control group. Then, we calculated descriptive statistics of the case and control group independently in the unmatching group. We applied conditional logistic regression in the case-control study through R package ‘survival’. Conditional logistic regression was chosen because of the imbalance between asthma patients and non-CRD patients. The statistically significant level was alpha equal to 0.05.

3. Result

A total of 48,640 COVID positive patients were included in this study, of which 5,557 patients were in the case group with one or more CRD, and 43,083 patients were in the control group without CRD. After we matched two groups with propensity-score matching method, there were 5,557 patients in case group and 11,114 patients in control group. Table 1 depicts the mortality rate of patients in different conditions. In the unmatched group, 1658 (3.9%) of non-CRD patients died due to COVID. The mortality rates due to COVID-19 in asthma patients were 2.9% (n=159). In the matched group, the overall mortality rate of asthma patients (2.9%) was significantly lower than that of patients without CRD (4.1%). Non-CRD patients in the matched group showed a higher mortality rate compared with non-CRD patients in unmatched group. This may be explained by bias reduction in matched cohorts by age, comorbidity score, and gender. Our analysis showed that asthma patients exhibited lower mortality. The two-proportions z-test shows the mortality rate in asthma patients has a significant difference from mortality rate in patients without CRD unmatched group, with p-value equal to 0.00029.

Table 1. Mortality Rate of patients with asthma and without CRDs.

	Death due to COVID	Alive	total	% Death Rate
With Asthma	159	5398	5557	2.9%
Without CRD (In Unmatched group)	1658	41398	43056	3.9%

Table 2 was the corresponding statistical summary table of the case and unmatched control groups. In the case group, there were 5557 (100%) asthma patients. Approximately 27.4% (n=1523) of people were identified as Hispanic or Latino, while in the control group, 27.0% (n=11616) were identified as Hispanic or Latino. In the control group, the percentage of the minority group has increased compared to the case group. White people accounted for the majority in the case group (62.4% (n=3467)) and still accounted for the majority in the control group, but the proportion was small, 58.8% (n=25312). Meanwhile, only 12.5% (n=693) of African Americans were in the case group, 13.2% (n=5701) of African Americans were in the control group. No matter in case group or control group, patients whose BMI is identity as overweight and obese are the majority group (respectively 78.8%, 71.5%). Underweight patients in both case and control groups only take 1.1 % 1.7%, respectively.

Table 2. Descriptive statistics of variables in case group and control group

		Patients only has asthma (Case) N=5557		Patients without CRD (control) in unmatched group N= 43083	
Variables Name		Numbers of patients	%	Numbers of patients	%
Ethnicity	Hispanic or Latino	1523	27.4%	11616	27.0%
	Not Hispanic or Latino	4034	72.6%	31467	73.0%
Race	White	3467	62.4%	25312	58.8%
	Black or African	693	12.5%	5701	13.2%
	Others	1397	25.1%	12070	28.0%
BMI	Underweight	62	1.1%	741	1.7%

	Normal weight	1115	20.1%	11546	26.8%
	Overweight and obesity	4380	78.8%	30796	71.5%
Asthma	Has Asthma	5557	100.0%	0	0.0%
Age status	18-35	1632	29.4%	12178	28.3%
	36-50	1422	25.6%	10401	24.1%
	51-75	2124	38.2%	16435	38.1%
	76+	379	6.8%	4069	9.4%
Comorbidity status	0-2	4086	73.5%	34376	79.8%
	3-4	967	17.4%	6652	15.4%
	5-6	321	5.8%	1356	3.1%
	7+	183	3.3%	699	1.6%
Gender (match)	MALE	1837	33.1%	20100	46.7%
	FEMALE	3720	66.9%	22983	53.3%

The conditional logistic regression results are presented in Table 3 as the odds ratio (OR) of each variable. Asthma patients compared to patients without CRD were significantly less likely (OR=0.7) to be deceased due to COVID-19 (p-value < 0.001). Race variable was as a significant confounder. Compared with Whites, African Americans were 1.88 times more likely to die from COVID-19 (p-value < 0.001). Compared with White, Others were 1.87 times more likely to be deceased from COVID-19 (p-value < 0.001). Hispanic or Latino patients were more likely to be deceased from COVID-19 when compared with patients who were non-Hispanic or Latino (OR=1.44, p-value < 0.023). Two BMI levels (underweight, overweight and obesity) didn't show a significant association with COVID-19 mortality. Compared with normal weight patients, underweight patients were 1.40 times more likely to die (p-value < 0.363) however this relationship didn't reach statistically significant level.

Table 3. Result of Conditional Logistic Regression to Identify Mortality Risk Factor Asthma

		Odds ratio	CI (Lower)	CI (Upper)	Pr(>|z|)
Patient	Do not have any CRD	1(ref)	_	_	_
	Asthma	0.7	0.58	0.86	0.001
Ethnicity	Not Hispanic or Latino	1(ref)	_	_	_
	Hispanic or Latino	1.44	1.05	1.96	0.023
Race	White	1(ref)	_	_	_
	Black or African American	1.88	1.39	2.55	< 0.001
	Others	1.87	1.37	2.55	< 0.001
BMI	Normal weight	1(ref)	_	_	_
	Underweight	1.4	0.68	2.9	0.363
	Overweight & obesity	1.01	0.78	1.32	0.912

4. Discussion

The conditional logistic regression results showed that asthma patients had a lower risk of death when compared with non-CRD patients within each stratum. That shown asthma would not increase the risk of death compared with non-CRD patients at a similar age,

gender comorbidity score. Previous studies demonstrated that asthma patients with COVID-19 had a high prevalence of comorbidities, including hypertension, heart disease, diabetes, and obesity [6]. Thus, not accounting for comorbidity status could have biased some studies and result in a conclusion that asthma increases mortality risk in COVID-19 patients [2]. Because of the feature selection and propensity matching in our study that included comorbidity score, this potential bias could have been alleviated. That could explain why our study showed that COVID-19 positive asthma patients had a lower death rate.

Propensity-score matching based on gender, comorbidity status, and age status in a case-control study allowed an adjustment for confounding and improved the study design. The imbalance between number of patients with asthma and patients without CRD in the original dataset was reduced by the 1:2 matching procedure. Each patient in the case group was matched with two patients in the control group that attenuated the effect of unmeasured confounders. Overall, there were 5,557 strata. Conditional logistics regression ensured the statistical power of subgroup analysis and interaction tests [7]. The estimated coefficients that were calculated by conditional logistic regression could only be interpreted within each stratum. Patients with other chronic diseases like diabetes may increase mortality rate. African Americans had a higher risk of death when tested positive for COVID-19 compared with Whites. Hispanic or Latino patients had also a higher mortality rate. BMI levels (overweight and underweight) did not reach a sufficient level of significance as a mortality risk factor in COVID-19 patients. Our study demonstrated the power of a nation-wide big data resource and utility of a secure data enclave for conducting big data analyses using case-control design.

5. Conclusion

When analyzing the COVID-19 mortality, underlying chronic respiratory conditions may play an essential role. Covid-19 positive asthma patients had a lower death rate than non-CRD patients. Using propensity-score matching in case-control design, we reduced the confounding effect of gender, age, and comorbidity with the. Race and ethnicity were important factors in the analysis of COVID-19 deaths.

References

[1] Dong E, Du H, Gardner L. An interactive web-based dashboard to track COVID-19 in real time. Lancet Inf Dis. 2020 May;533-534.

[2] Shi T, Pan J, Vasileiou E, Robertson C, Sheikh A. Public Health Scotland and the EAVE II Collaborators. Risk of serious COVID-19 outcomes among adults with asthma in Scotland: a national incident cohort study. Lancet Respir Med. 2022 Jan;S2213-2600(21)00543-9.

[3] Haendel MA, Chute CG, Bennett TD, et al. The National COVID Cohort Collaborative (N3C): Rationale, design, infrastructure, and deployment. Journal of the American Medical Informatics Association: JAMIA vol. 2021 Mar;427-443.

[4] Ho CH, Chen YC, Chu CC, Wang JJ, Liao KM. Age-adjusted Charlson comorbidity score is associated with the risk of empyema in patients with COPD. Medicine (Baltimore). 2017 Sep;96(36), e8040.

[5] Cui W, Cabrera M, Finkelstein J. Latent COVID-19 Clusters in Patients with Chronic Respiratory Conditions. Stud Health Technol Inform. 2020 Nov;275:32-36.

[6] Morales D, Ostropolets A, Lai L, et al. Characteristics and outcomes of COVID-19 patients with and without asthma from the United States, South Korea, and Europe. J Asthma. 2022 Jan:1-14.

[7] Kuo CL, Duan Y, Grady J. Unconditional or Conditional Logistic Regression Model for Age-Matched Case-Control Data? Front Public Health. 2018 Mar;6:57.

Challenges of Trustable AI and Added-Value on Health
B. Séroussi et al. (Eds.)

doi:10.3233/SHTI220474

Proposal of Semantic Annotation for German Metadata Using Bidirectional Recurrent Neural Networks

Hannes ULRICH[a,1], Hristina UZUNOVA[b], Heinz HANDELS[b,c] and Josef INGENERF[a,c]
[a] *IT Center for Clinical Research (ITCR-L), University of Lübeck, Germany*
[b] *German Research Center for Artificial Intelligence, Lübeck, Germany*
[c] *Institute of Medical Informatics, University of Lübeck, Germany*

Abstract. The distributed nature of our digital healthcare and the rapid emergence of new data sources prevents a compelling overview and the joint use of new data. Data integration, e.g., with metadata and semantic annotations, is expected to overcome this challenge. In this paper, we present an approach to predict UMLS codes to given German metadata using recurrent neural networks. The augmentation of the training dataset using the Medical Subject Headings (MeSH), particularly the German translations, also improved the model accuracy. The model demonstrates robust performance with 75% accuracy and aims to show that increasingly sophisticated machine learning tools can already play a significant role in data integration.

Keywords. Metadata, Unified Medical Language System, Deep Learning

1. Introduction

The digital transformation is progressively changing our healthcare system and its disciplines. The rapid pace of digitization is also creating more new clinical data sources that need to be analyzed and integrated to be used together. This is of enormous importance for patient care because a data fusion of all sources allows a comprehensive, holistic overview. The fragmentation of digital healthcare into many proprietary individual systems and data formats makes the desired overview difficult and slows down technical innovations. The clinical data integration shall gap this fragmentation and is an essential foundation for further data processing. One suitable tool in this context is metadata [1], which is able to describe the diverse characteristics of information objects precisely. In addition to content and administrative information, they also suitably depict the structure and - with the help of annotations - the semantics. The semantic coding enables a better understanding of the described data but is only useful if the annotation is carried out extensively and is sustainable. It must be ensured that codes or the coding system also fit the described content of the metadata and that the annotations are carried out consistently [2]. In contrast, inconsistent annotations degrade

[1]Corresponding Author, Hannes Ulrich, IT Center for Clinical Research, Lübeck; Telephone: +49 (0) 451 - 3101 5607; Fax: +49 451 3101 5604; E-mail: h.ulrich@uni-luebeck.de

data quality and findability. Manual annotation is time and resource-consuming, so machine support is desirable [3]. However, this support should not simply propose codes but should incorporate previous annotations to ensure consistent annotation. The already annotated datasets provide implicit knowledge about previous annotations that should be used meaningfully. Neural networks are the state of the art to exploit this implicit knowledge and to support annotators. To our knowledge, there are no comparable approaches in the literature that describe a predictive model of semantic annotations for given German metadata.

2. Methods

The proposed approach uses a trained neural network-based model to propose semantic codes corresponding to a given metadata item. In the following, the used dataset and its preprocessing steps are described, as well as the architecture of the neural network.

2.1. Dataset and Preprocessing

The dataset originates from the MDM Portal [4], which collects medical data models, mostly electronic Case Report Forms (eCRFs) from clinical research form data. The MDM currently contains ca. 24.000 forms with 500.000 metadata items in 53 languages. The special characteristic of the MDM is the manual annotations of the forms using Unified Medical Lanugage System (UMLS) codes [5], which is done by medical experts. Since our work focuses on the annotation of German metadata, only forms in German were considered. A total of approx. 150.000 annotated German items were available, which described the question groups, the questions, and structured answer options. Before the dataset was used for training, it had to be cleaned, and then additional data was augmented. The cleaning removed samples with deprecated annotations and codes with less than 50 occurrences were sorted out. To increase the amount of training data and increase robustness, the dataset was augmented using the Medical Subject Headings (MeSH) [6]. The MeSH thesaurus which is mainly used for indexing and retrieval of literature appearing in MEDLINE/PubMed, provides German translations of subject headings. As this German version is included within the UMLS metathesaurus, the selection of codes leads to additional 11.700 samples.

```
C0421448;name des patienten
C0421448;patienten-nachname
C0421448;name
C0421448:nachname
```

Figure 1. The input data are pairs of a UMLS code and the corresponding phrase, for example, the patient surname (in German).

2.2. Network Architecture and Training

The expected input is a metadata definition, a sequence of words representing the display text of the metadata item, e.g., “The name of the patient”. Sequential deep learning models, particularly recurrent neural networks (RNNs) and bidirectional long short-term memories (BiLSTMs), have shown robust performance on tasks which require the

encoding of short sequences of words [7, 8]. In order to predict the most probable semantic annotation, the input sequences are passed through trough the layers of a LSTM-based network responsible for transforming the definition into a computable representation, learning the sequential correspondences within the words of the definition, and predicting the most probable semantic code. The network architecture is shown in Figure 1. At first, the sequences are split into word tokens, and each unique word is represented through an index. These indices are passed into an embedding layer that transforms each word into a vector representation. These representations are passed into a two-tier BiLSTM layer and then into the last classification layer. The output is mapped to possible classes representing UMLS codes.

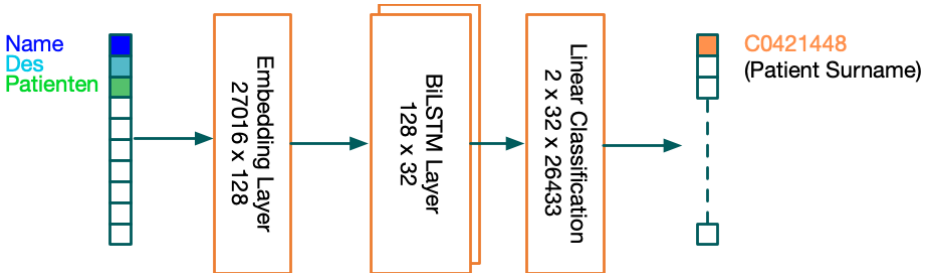


Figure 1. The trained network is built of three consecutive layers. The embedding layers encoded the words into a computable vector representation, which will be used in the next layer. In the two-cell bidirectional LSTM, the word sequences ("Name des Patienten") are learned and the resulting output is mapped on the labels in the linear layer.

The network was implemented in Python using Pytorch and is available on Github [9]. The embedding layer was parametrized with 26433 unique words and 128 dimensions, the two-tier BiLSTM with 32 hidden dimensions. The input and output dimensions of the linear classifier were set to 25992 representing every possible code in the datasets and with two times of the hidden dimensions of the BiLSTM since the output dimension is double due to the fact of using a BiLSTM.

We use an 80/20 training/test split. 20% of the training data are then used for validation during the training to determine the optimal number of training epochs. An Adam-optimizer with a learning rate of 0.01 is used. The training was carried out for 250 epochs with a batch size of 5000. A cross-entropy loss function is used weighted by a factor $w = l/l_{max}$ where l is the current length of the phrase and l_{max} is the maximal phrase length in the training set. The intuition behind this weighting is to give each word the same relative importance to the loss, such that short and one-word phrases do not overweight long phrases.

3. Results

Different variants were trained to identify the best network configuration. The depth of the BiLSTM layer was varied to achieve the best accuracy on the given data set. The best configuration was trained for another experiment with German GloVe embeddings [10] instead of the inbuilt embedding layer. These embeddings were trained on large text bodies and can better detect synonyms in the data. However, the accuracy was lower compared to the learned embeddings. Then the configured networks were trained with and without the augmented data. The results of the experiments are shown in **Table 2,**

where T*k* indicates that the predicted label was contained in the top *k* most probable predictions. The best results were achieved by a two-tier BiLSTM closely followed by the one-tier architecture. Furthermore, the results underline the importance of augmented data and its potential to achieve even higher accuracy. Overall, a fair accuracy of ca. 67% could be achieved, where an even higher accuracy of 75% can be observed when the ten best predictions are considered.

Table 1. The table shows the first three predictions for the given set with the probabilities in brackets.

Phrase	1. Prediction	2. Prediction	3. Prediction
Name des Patienten (name of the patient)	Patient surname (25.04%)	Medication name (18.32%)	Patient forename (17.45%)
Krebs der Niere (cancer of the kidndy)	Kidney cancer (24.84%)	Subject Diary (17.16%)	Malignant neoplasm of kidney (16.12%)
Krebs der Leber (cancer of the liver)	Malignant Placental Neoplasm (21.20%)	Secondary malignant neoplasm of liver (18.50%)	Liver reconstruction (17.72 %)
Blut (blood)	Blood (19.38%)	Blood in Urine (17.81%)	Coagulation Process (17.21%)

Table 2. The table presents the results of the conducted experiments. We trained models with four different configurations: three different layer depths of the LSTM layer and additionally with pre-trained GloVe embeddings for the best model configuration. The experiments were conducted in each configuration with and without the augmented data. Here, T1, T3, T5, T10 signifies whether the ground truth class was within the first, first three, first five, or first ten most likely predictions. The best results were achieved by the two-tier BiLSTM.

Experiment	T1 Acc.	T3 Acc.	T5 Acc.	T10 Acc.
BiLSTM w/o Augmentation	63.77	70.16	71.39	72.69
BiLSTM w/ Augmentation	67.04	72.66	73.9	75.44
2BiLSTM w/o Augmentation	64.63	70.95	71.2	73.27
2BiLSTM w/ Augmentation	**67.27**	**72.85**	**74.23**	**75.63**
2BiLSTM w/o Glove w/o Aug.	60.09	66.16	67.34	68.76
2BiLSTM w/ Glove w/o Aug.	65.09	70.56	72	73.64
3BiLSTM w/o Augmentation	63.46	69.17	70.42	71.53
3BiLSTM w/ Augmentation	66.23	71.45	72.66	74.03

4. Discussion

The specificity of the proposed model resides in the processing and semantic annotation of German metadata - to our knowledge, there is no comparable model. The accuracy of the predictions is sound, although there is potential for improvement. In addition, only a fraction of all UMLS Codes were included in the data set, so not all concepts can be predicted. For example, the concept “liver cancer” was not in the dataset, so the network recognized liver and cancer distinctly, but the joining concept was unknown, as seen in **Table 1**. The use of augmented data showed an improvement in accuracy in all experiments. Therefore, further augmentation sources should be used in subsequent studies to enhance the training dataset. The use of pre-trained GloVe embeddings should allow a better understanding of synonyms. Nevertheless, the overall accuracy was worse. One possible explanation is that most keywords in the phrases are specialized medical vocabularies and are often not included in models pre-trained on general text corpora. Future work will consider and further refine training on the target GloVe embedding dataset. Another possibility is the use of newer attention-based models such as

Transformer, but there is currently no biomedical Transformer model for the German language. One inaccuracy in general remains for the process of data integration using the trained model: the predictions are based on previous annotations. Misclassifications then carry over into subsequent processing steps. However, the MDM portal can also draw on great prior work in terms of consistency and interrater variability [11], so that the use of the model can be recommended.

5. Conclusions

This work is intended as a proof-of-concept to show that increasingly performant machine learning tools can already play an important role in data integration - effectively a stage before the initial provision of curated research data. The proposed model can usefully support annotators to enable new datasets for secondary research and hopes to be an impetus for future work in the area, such as integrating the UMLS graph knowledge into the network.

Acknowledgments

The authors would like to thank the MDM team for their long-standing and great work of providing and annotating medical forms. This work was supported by the German Research Foundation (Deutsche Forschungsgemeinschaft) DFG grants IN 50/3-2.

References

[1] Ulrich H, Kock-Schoppenhauer AK, Deppenwiese N, Gött R, Kern J, Lablans M, Majeed RW, Stöhr MR, Stausberg J, Varghese J, Dugas M, Ingenerf J. Understanding the Nature of Metadata: Systematic Review. J Med Internet Res. 2022 Jan 11;24(1):e25440. doi: 10.2196/25440. PMID: 35014967; PMCID: PMC8790684.

[2] Cardoso SD, Pruski C, Da Silveira M, Lin Y-C, Groß A, Rahm E, et al., Leveraging the impact of ontology evolution on semantic annotations, in: Springer, 2016: pp. 68–82.

[3] Ulrich H, Kock-Schoppenhauer AK, Andersen B, Ingenerf J. Analysis of Annotated Data Models for Improving Data Quality. Stud Health Technol Inform. 2017;243:190-194. PMID: 28883198..

[4] Dugas M, Neuhaus P, Meidt A, Doods J, Storck M, Bruland P, Varghese J. Portal of medical data models: information infrastructure for medical research and healthcare. Database (Oxford). 2016 Feb 11;2016:bav121. doi: 10.1093/database/bav121. PMID: 26868052; PMCID: PMC4750548.

[5] Bodenreider O, The unified medical language system (UMLS): integrating biomedical terminology, Nucleic Acids Research. 32 (2004) D267–D270.

[6] ZB MED – Information Centre for Life Sciences, German MeSH, ZB MED - Informationszentrum Lebenswissenschaften. (2022). https://www.zbmed.de/en/open-science/terminologies/german-mesh/ (accessed January 21, 2022).

[7] Xu K, Zhou Z, Hao T, and Liu W, A bidirectional LSTM and conditional random fields approach to medical named entity recognition, in: Springer, 2017: pp. 355–365.

[8] de Buy Wenniger GM, van Dongen T, Aedmaa E, Kruitbosch HT, Valentijn EA, and Schomaker L, Structure-tags improve text classification for scholarly document quality prediction, in: 2020: pp. 158–167.

[9] Ulrich H, mentored, Zenodo, 2022. doi:10.5281/zenodo.5897287.

[10] Pennington J, Socher R, and Manning CD, Glove: Global vectors for word representation, in: 2014: pp. 1532–1543.

[11] Varghese J, Sandmann S, Dugas M. Web-Based Information Infrastructure Increases the Interrater Reliability of Medical Coders: Quasi-Experimental Study. J Med Internet Res. 2018 Oct 15;20(10):e274. doi: 10.2196/jmir.9644. PMID: 30322834; PMCID: PMC6231825.

Challenges of Trustable AI and Added-Value on Health
B. Séroussi et al. (Eds.)

doi:10.3233/SHTI220475

TerminoDiff – Detecting Semantic Differences in HL7 FHIR CodeSystems

Joshua WIEDEKOPF[a,b,1], Cora DRENKHAHN[a,b], Lorenz ROSENAU[a,b], Hannes ULRICH[a,b], Ann-Kristin KOCK-SCHOPPENHAUER[b] and Josef INGENERF[a,b]
[a] *Institute of Medical Informatics, University of Lübeck, Lübeck, Germany*
[b] *IT Center for Clinical Research, University of Lübeck, Lübeck, Germany*

Abstract. While HL7 FHIR and its terminology package have seen a rapid uptake by the research community, in no small part due to the wide availability of tooling and resources, there are some areas where tool availability is still lacking. In particular, the comparison of terminological resources, which supports the work of terminologists and implementers alike, has not yet been sufficiently addressed. Hence, we present TerminoDiff, an application to semantically compare FHIR R4 CodeSystem resources. Our tool considers differences across all levels required, i.e. metadata and concept differences, as well as differences in the edge graph, and surfaces them in a visually digestible fashion.

Keywords. Health Information Exchange; Terminology as Topic; Vocabulary; HL7 FHIR

1. Introduction

1.1. Background

The HL7 *Fast Healthcare Interoperability Resources* (FHIR) specification has seen an enthusiastic adoption by the Medical Informatics research community for the interoperable storage and exchange of medical information [1,2]. The enthusiastic uptake of this very recent specification is substantially due to the strong focus by the standards developers on the availability of tooling [1].

Parallel to the establishment of FHIR in healthcare IT, there is an increasing drive to incorporate standard terminology throughout the ecosystem, in order to maintain semantic interoperability of healthcare data across systems, institutional and even national boundaries. The terminological systems ICD-10, ICD-11, SNOMED CT and LOINC, among many others, play an important part in this development [1,3]. These terminological systems are maintained independently of the FHIR specification, but can be queried via the mechanism of FHIR terminology servers [3].

However, in real-world systems, standardized terminology has yet to fully supplant proprietary coding systems. Hence, the creation of interoperable terminology resources is an important part of ensuring healthcare interoperability on a whole; for which the

[1] Corresponding Author, Joshua Wiedekopf, Institute of Medical Informatics, University of Lübeck, Ratzeburger Allee 160, 23564 Lübeck, Germany; E-Mail: j.wiedekopf@uni-luebeck.de

HL7 FHIR Terminology Module provides a pathway that has seen adoption even outside of FHIR-compliant systems, like within the openEHR specification [4]. Additionally, to make use of the terminological knowledge contained in coding systems like ICD-10-GM, being the federally mandated classification for diagnoses within Germany, those resources need to be converted to FHIR directly [5].

There is tooling available both for the automatic and for the manual creation of terminology resources in FHIR [3,5]; however, little exists in the area of continuing maintenance of these resources. The development of these tools not only aids developers of terminologies, classifications and other artefacts, but also helps implementers of HL7 FHIR in migrating from one version to another.

While existing terminological systems generally provide some aids for transitioning to newer versions, such as history relationships in SNOMED CT or delta tables for ICD-10, maintenance of FHIR resources is not standardized, necessitating tooling support. Hence, we present a method and implementation for the computation and visualization of the differences between versions of the same HL7 FHIR CodeSystem.

1.2. Related Work

We have carried out a scoping literature review in order to get an overview of established algorithms in this field, searching the digital libraries *PubMed*, *Scopus* and *Springer Link* with the search string *("fhir" OR "ontology" OR "rdf" OR "terminology") AND ("diff" OR "change" OR "difference" OR "version")*.

In total, seven studies were considered applicable to the present work; more detail on this review is available online in our GitHub repository[2]. Of these seven studies, none considered the maintenance of any kind of HL7 FHIR resources (including non-terminological resources like *Patient*); all studies were examining the problem from the view of formal ontologies and the semantic web.

The existing work in this closely related field [6,7] illustrates a clear need for the development of a difference computation. Additionally, the existing approaches generally do not provide a graphical (as in graph-based) view of the changes in connection between the versions, which we deem crucial for obtaining an understanding of relation between individual changes.

2. Methods

In HL7 FHIR, there are three kinds of terminological resources, namely *CodeSystem*, *ValueSet* and *ConceptMap*. We have focused on *CodeSystem* in this work, which is the resource type used for declaring the existence of a coding system with its associated metadata, and (generally) the concepts contained within that coding system with their associated relationships. There can be more than one kind of relationship, and polyhierarchical associations (where one concept may have more than one parent concept) are supported.

One example of a CodeSystem resource is one describing the currently 884 concepts of the OncoTree cancer classification, which are related in mono-hierarchical *parent*-relationships [8].

[2] https://itcr-uni-luebeck.github.io/TerminoDiff/SLR

Table 1. Levels of differences we identified for FHIR R4 CodeSystem resources, and proposed resolution strategy

Level	Aspect	Example	Resolution strategy
1	Metadata-level		Presentation as a table in the GUI
1.1	Simple differences	*title*, *name, version*	String comparisons
1.2	Differences within lists	*identifier*, *language*	(keyed) difference lists, e.g. by *language.code*
2	Concept-level		Presentation as a table in the GUI
2.1	Simple differences	*display*, *definition*	String comparisons
2.2	Differences within lists	*property, designation*	(keyed) difference lists, e.g. by *property.code*
2.3	Unilaterality of concepts	Deletions and additions of codes / concepts across versions	Surfacing in the table with dedicated filter and highlighting
3	Edge differences	Changes to properties linking concepts, i.e., *parent*	Creation and visualization of a difference graph

For obtaining a meaningful comparison of two CodeSystem resources, multiple levels must be considered, utilizing different strategies; for example, the metadata of the resource presents different challenges than the concept relationships. The levels of difference we identified for FHIR CodeSystem resources are provided in Table 1.

To illustrate the third level of differences, consider Figure 1, illustrating a difference graph for a fictitious pair of code graphs. While FHIR R4 allows at least three different approaches to specifying a *parent*-*child*-relationship (a property with code *parent*, one with code *child*, and the *concept* element within *concept*), we consider these to be semantically equivalent and reduce them to a canonical *parent* property. Our approach considers both the possibility of polyhierarchical relationships, where a concept may have more than one parent, as well as the possibility for multiple types of edges (such as *related-to* in this example).

3. Results

3.1. Implementation

The proposed software was implemented using the *Kotlin* programming language as a desktop application with a graphical user interface, utilizing the *Compose Desktop* toolkit. Our graph-based algorithms utilize the JGraphT library [9]

At the foundation of our implementation lie several components that build a difference model from the two provided CodeSystem resources. The user chooses one of these to be the *left* resource, the other the *right*, taking a view that is consistent with the side-by-side view commonly provided by generic comparison programs. Generally, the *right* resource should be chosen as a newer version of the *left* resource. For working with FHIR resources, we utilize the well-known *HAPI FHIR* library.

As the difference model must be constructed across several levels, as illustrated in Table 1, our engine also is split across multiple components. Metadata differences at Level 1 are shown to the user as a table; the order of items is chosen to be consistent to the FHIR specification of the CodeSystem resource.

Another component computes differences on the lists of concepts provided in each CodeSystem resource, surfacing values that are only referenced in one resource, as well as differences that occur in concepts referenced in each side.

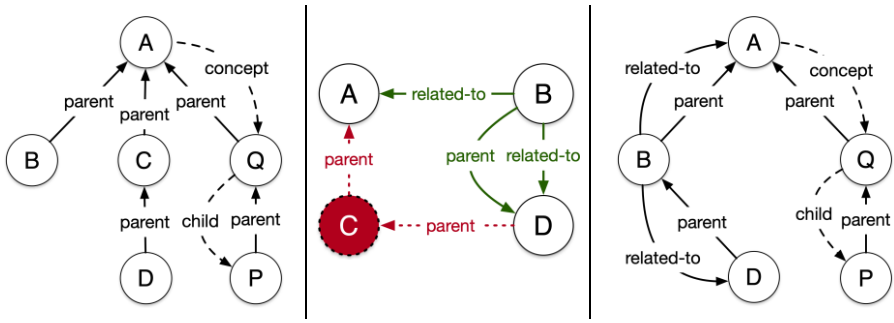


Figure 1. Two fictitious code graphs representing two versions of a code system (left, right), and the difference graph (middle). In the difference graph, vertices and edges that only occur in the left code graph are shaded red dotted lines, those only in the right code graph are shaded green with solid edges. Dashed edges reduced to a *parent* relationship in this approach.

The third component renders a difference graph for the user. We compute a graph representing the additions and deletions of concepts and edges, illustrating the changes between the different versions. We do this in the same fashion as shown in Figure 1, allowing for a visual assessment of the changes within the CodeSystem hierarchy.

We make the source code of our program freely available via GitHub[3] and Zenodo[4] under the terms of the GPL 3.0 license.

3.2. Evaluation

The implementation was tested both with constructed examples, where the differences in metadata and/or concept relationships were well-known, as well as with real-world examples that were available in HL7 FHIR. For the latter, we utilized a number of examples, such as resources that have been specified in the context of the *GECCO* dataset by the German *Network University Medicine* [10], different versions of the *ICD-10-GM* classification, and the aforementioned *OncoTree* cancer classification [8].

Supporting our claim that our tool can improve the process of maintaining terminological artefacts, we were able to spot an omission in the OncoTree release notes (since corrected pursuant to our report), whereby a concept that has been introduced in version 2021-11-02 has not been referenced in the release notes.

As our software is still in active development, there has not yet been any formal evaluation of the tool with users not involved in software development; however, we are currently planning a survey among experts in terminology creation and maintenance that are active in the Medical Informatics Initiative in Germany and elsewhere.

4. Discussion

We believe that our implementation and framework can aid terminological authors in their day-to-day work of creation and maintenance, such that the tool can lead to greater acceptance and adoption of HL7 FHIR terminology and terminology servers at large

[3] GitHub repository: https://github.com/itcr-uni-luebeck/TerminoDiff
[4] Zenodo DOI: 10.5281/zenodo.5898267

scale. We believe this goal to be of supreme importance to the broader Medical Informatics research community to be able to ensure semantic interoperability across systems.

Furthermore, our approach could likely be amended to also consider other types of HL7 FHIR resources, such as the other two terminological resource types *ValueSet* and *ConceptMap*, but also other definitional artefacts, such as *StructureDefinition* (used for describing profiles of FHIR instance data), the maintenance of which we believe to benefit from additional tooling support as well.

5. Conclusion

The availability of user-friendly tooling is an important factor in the acceptance of standard in the industry. We have developed an application that tackles a challenge associated with developing new versions of terminological resources, as well as adopting these in applications. Doing so, we provide a tool that can help to increase the degree of semantic interoperability across the healthcare information landscape. In this fashion, we aid the transition from local in-house terminology to interoperable specifications, which is called for by the research community and political decision-makers alike.

Acknowledgements

This work was supported by the HiGHmed project within the Medical Informatics Initiative, funded by the Federal Ministry of Education and Research (grant number 01ZZ1802Z).

References

[1] Benson T, Grieve G. Principles of Health Interoperability. Third edition. Springer International Publishing; 2016. https://doi.org/10.1007/978-3-319-30370-3.

[2] Lehne M, Luijten S, Vom Felde Genannt Imbusch P, Thun S. The Use of FHIR in Digital Health - A Review of the Scientific Literature. Stud Health Technol Inform 2019;267:52–8. https://doi.org/10.3233/SHTI190805.

[3] Metke-Jimenez A, Steel J, Hansen D, Lawley M. Ontoserver: a syndicated terminology server. J Biomed Semant 2018;9:24. https://doi.org/10.1186/s13326-018-0191-z.

[4] openEHR. Archetype Query Language (AQL) 2021. https://specifications.openehr.org/releases/QUERY/latest/AQL.html (accessed January 11, 2022).

[5] Wiedekopf J, Drenkhahn C, Ulrich H, Kock-Schoppenhauer A-K, Ingenerf J. Providing ART-DECOR ValueSets via FHIR Terminology Servers - A Technical Report. Stud Health Technol Inform 2021;283:127–35. https://doi.org/10.3233/SHTI210550.

[6] Hartung M, Groß A, Rahm E. COnto–Diff: generation of complex evolution mappings for life science ontologies. J Biomed Inform 2013;46:15–32. https://doi.org/10.1016/j.jbi.2012.04.009.

[7] Klein M, Fensel D, Kiryakov A, Ognyanov D. Ontology Versioning and Change Detection on the Web. Knowl. Eng. Knowl. Manag. Ontol. Semantic Web, Springer Berlin Heidelberg; 2002, p. 197–212. https://doi.org/10.1007/3-540-45810-7_20.

[8] Kundra R, Zhang H, Sheridan R, Sirintrapun SJ, Wang A, Ochoa A, et al. OncoTree: A Cancer Classification System for Precision Oncology. JCO Clin Cancer Inform 2021:221–30. https://doi.org/10.1200/CCI.20.00108.

[9] Michail D, Kinable J, Naveh B, Sichi JV. JGraphT—A Java Library for Graph Data Structures and Algorithms. ACM Trans Math Softw 2020;46.

[10] Sass J, Bartschke A, Lehne M, Essenwanger A, Rinaldi E, Rudolph S, et al. The German Corona Consensus Dataset (GECCO): a standardized dataset for COVID-19 research in university medicine and beyond. BMC Med Inform Decis Mak 2020;20:341. https://doi.org/10.1186/s12911-020-01374-w.

Challenges of Trustable AI and Added-Value on Health
B. Séroussi et al. (Eds.)

doi:10.3233/SHTI220476

FAIRifying a Quality Registry Using OMOP CDM: Challenges and Solutions

Daniel PUTTMANN[a,b,1], Nicolette DE KEIZER[a,b], Ronald CORNET[b], Eric VAN DER ZWAN[a,b] and Ferishta BAKHSHI-RAIEZ[a,b]
[a] *The National Intensive Care Evaluation (NICE) registry, The Netherlands*
[b] *Department of Medical Informatics, Amsterdam Public Health Research Institute, Amsterdam UMC, University of Amsterdam, The Netherlands*

Abstract. The need for health data to be internationally Findable, Accessible, Interoperable and Reusable (FAIR) and thereby support integrative analysis with other datasets has become crystal clear in the ongoing pandemic. The Dutch National Intensive Care Evaluation (NICE) quality registry adopted the Observational Medical Outcomes Partnership Common Database Model (OMOP CDM) to achieve a FAIR database. In the process of adopting the OMOP CDM, many modeling, technical, and communication challenges needed to be solved. Through communication with the OMOP CDM implementation community, previously done research and trial-and-error we found solutions that we believe can help other healthcare institutions, especially ICU quality registries, FAIRify their databases.

Keywords. OMOP CDM, ETL Process, Quality Registry, OHDSI

1. Introduction

Though cooperation has always been important in research, the COVID-19 pandemic showed just how important it is to federate local databases for large-scale analysis efficiently. Both for research and healthcare, large amounts of data were and are still needed to combat outbreaks, to develop novel treatments and to gain insight into the new disease. However, time is short, and data is fragmented.

In general, health data is recorded in hospitals in a flexible, unstructured format or in a proprietary structured format. Databases are modeled with a clear, locally implemented goal in mind. This makes the data convenient for its original use case, but difficult to reuse for other purposes [1]. The databases of healthcare institutions are often modeled in this way, making them poorly interoperable.

Since the ’90s, researchers and standard development organizations have developed models to standardize healthcare databases. As a result, communities have grown from the different standardization paradigms, one being the Observational Health Data Sciences and Informatics (OHDSI) program [2]. The OHDSI program is an international collaborative aiming to analyze and perform research on large amounts of health data. It developed the Observational Medical Outcomes Partnership Common Database Model (OMOP CDM) to facilitate interoperability between databases. Interoperability here

[1] Corresponding Author, Daniel Puttmann; E-mail: d.p.puttmann@amsterdamumc.nl.

means that databases can be used to collaborate and combine data by using the same format, with enough metadata to make unfamiliar data sources findable and reusable. OMOP CDM can be used to work towards a Findable, Accessible, Interoperable and Reusable (FAIR) data infrastructure [3]. By making health data FAIR, medical research could make better use of technical advancements such as machine learning and the data sharing across the internet. In Europe, the European Health Data & Evidence Network (EHDEN) was launched to facilitate international, large-scale research with real-world health data. They have been collaborating with OHDSI since their inception in 2018.

The National Intensive Care Evaluation (NICE) registry is a quality registry developed by Dutch ICUs to monitor and improve the quality of care by learning from other ICUs. Despite some cross-registry projects, international collaboration with other ICU registries such as the Australian ANZICS-CORE or the Sri Lankan Critical Care Asia (CCA) is hampered by use of local definitions and lack of interoperability [4].

By adopting OMOP CDM and by joining the EHDEN-OHDSI community, NICE aimed to make its data FAIR and accessible for international researchers and align itself with other ICU registries adapting the OMOP, such as the CCA. With this, we let the ICU community benefit from larger datasets.

The aim of this paper is firstly to describe the different steps of the Extract, Transform and Load (ETL) process to make the NICE database FAIR by using OMOP CDM and OHDSI tools and secondly to investigate the challenges that were faced and the solutions that were found during the ETL process.

2. Methods

We performed the ETL on six tables from the NICE database [5]. In total there were 158 data elements adapted in English to the OMOP CDM. The FAIRification process was divided into four steps (see Figure 1). To complete the various steps, we used the Book of OHDSI, EHDEN academy's ETL course, OHDSI's GitHub wiki, and the OHDSI tools [6-9].

In the first step, we used the White Rabbit, Rabbit-In-A-Hat and Usagi OHDSI tools to analyze and map the data to the OMOP CDM and its underlying standard vocabularies.

The second step was the Extract, Transform and Load (ETL) process, which we performed using twelve Microsoft SQL Server (MSSQL) scripts. The database needed to be configured with empty tables representing the OMOP CDM and constraints were set for primary and foreign key rules. The OMOP CDM tables were populated based on the results from step one.

In the third step, we validated the ETL by performing two quality checks, for each of OHDSI's data quality management tools. These were the Data Quality Dashboard and the ACHILLES Heel. We evaluated the error messages and deleted the offending rows from the database where appropriate, per OHDSI's recommendation. Lastly, we implemented the Atlas research platform to allow for data analysis on our OMOP CDM database and we published a link to our database on the EHDEN data portal for researchers to contact us from.

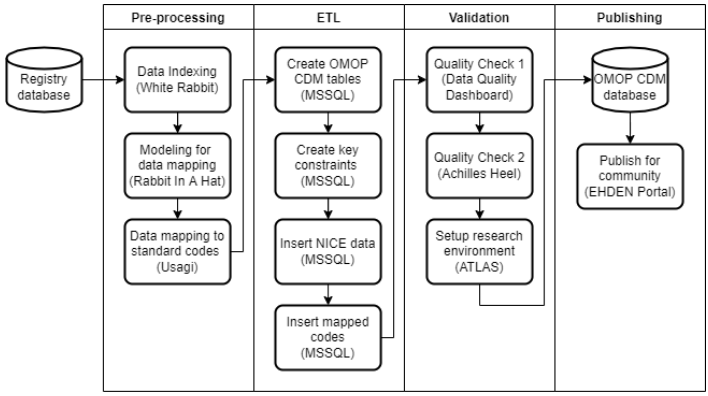


Figure 1. A flowchart of the process from the registry database to the OMOP CDM database. The tools used are mentioned in parentheses. All tools except MSSQL are developed by OHDSI.

3. Results

In this section we describe the challenges and the solutions, categorized in five main topics, namely mapping problems, differences in unit of observation, differences in structure of the database, security concerns and technical restrictions.

Mapping problems: we were unable to map 50 out of 158 data elements because no standardized code was available to fully reflect the concept's meaning. Five of these data elements and their values were part of specialized code sets such as for ICUs commonly used APACHE IV reasons for admission classification [10]. Finding a solution to this was challenging due to the lack of documentation on what to do when no suitable code was available. On OHDSI forums we found two solutions based on a conference poster [11]:

- Use a lookup table to insert custom codes in available tables after inserting source codes or headers
- Insert custom codes at the end of the imported OHDSI standard codes and create relationships such as 'maps to' in the concept relationship table

We implemented a hybrid form where each previously unmapped data element and value of the specialized code sets were assigned custom codes and code relationships. They were then inserted in the right OMOP CDM tables using a lookup table.

Unit of observation: the NICE database was centered around ICU visits, not around patients. This was an issue when generating primary keys for OMOP CDM tables, which made linking the OMOP CDM tables difficult. This was solved by generating unique keys for each part of the NICE database's composite primary key, used to designate unique ICU visits. This key included an encrypted patient id, a hospital admission number, and an ICU admission number. Respectively, for the person, visit occurrence and visit detail tables we created three lookup tables that generated primary keys from different combinations of the NICE primary keys and some patient characteristics if necessary.

Structure of the database: the NICE tables were in a wide format while the OMOP CDM tables were in a long format. Every data item name was a header in the NICE table, while the names were put in a single column in the OMOP CDM. This was eventually solved by creating a temporary table with the desired data elements before each insertion

and using the 'UNPIVOT' SQL operator. It takes multiple columns and collates them into two columns: one column for the header and another for the value belonging to that header. If the data type of the source value did not match with the data type of the target column, we converted the source value to that data type. We then inserted the data into the target table using the unpivoted temporary table.

Security concerns and technical restrictions: the use of the university infrastructure with all its constraints lead to delays in the ETL process. The data quality tools validating the database required R scripts to run, which required specific versions of certain R and java packages to be installed. These installations needed to be approved by the IT department due to protocol. Furthermore, we chose to implement ATLAS via a Docker container because of the reported difficulty in implementation. The ATLAS Docker container is a package containing ATLAS and all its dependencies. The Docker also included a major dependency called WebAPI that connects the OMOP CDM database server to its own server. We were reluctant to install the WebAPI on the NICE database as it was not intended to be connected to the internet. These concerns were a result of a misnomer. The term WebAPI usually refers to a way for external applications to interface with a data source. Although the WebAPI java application has the possibility to interface with such applications, it can also be installed and used locally. To run ATLAS, 104 new tables were needed to store results in. There was no script on the OHDSI Github that could generate these tables, but we could acquire a script from the OHDSI community.

The last *technical hurdle* was that, when creating a cohort to do analysis on, we were not able to use the function to create inclusion criteria due to key duplication errors. After consulting the OHDSI implementation forums we found a solution that had us delete primary keys in the ATLAS cohort inclusion tables [12]. The issue was that these tables wrongly had primary keys assigned to them.

4. Discussion and Conclusions

By implementing OMOP CDM, NICE can now use the powerful ATLAS and R toolkit of OHDSI to do research and share their data with researchers across the globe. Our future research should focus on the capabilities of the new database and the further standardization of the data elements.

OHDSI's ETL process had from an implementer's perspective a very low technical cohesion. Meaning that in each step new applications had to be installed, new programming languages had to be used and new documentation had to be searched for on different platforms. We suggest OHDSI to focus on the coherence of their ETL process to improve the rate of adaptation. More healthcare institutions will be able to adopt OMOP CDM if the materials and documentation are not as fragmented as they are now, and if software works without having to install many specific dependencies or writing additional code.

To solve documentation fragmentation, the EHDEN-OHDSI collaboration developed the Book of OHDSI and EHDEN academy [6,7]. However, they conflate theory and practice. Ideally, the Book of OHDSI would be a main source for background information and the EHDEN academy a source for implementation guides. Moreover, OHDSI's tutorials can be followed with an Amazon virtual machine called OHDSI-in-a-box, which simulates a database and all ETL tooling to lead users through the process using fake data. However, to our knowledge there are no alternatives for implementers without an Amazon paid subscription.

Technical difficulties could be solved with the creation of Docker containers for every application. The OHDSI community has already started a project on this called BROADSEA [13]. If OHDSI's Docker containers worked out of the box, it would not only accelerate the process, but also make the task less daunting.

We expect this paper to be of use to most other health organizations that want to FAIRify their database using the OMOP CDM. The challenges we faced had very common causes: many medical databases have a different unit of observation, a different structure, or data elements with specific, hard to standardize meanings. When mapping data, the list of data elements should be inspected for unmappable codes. These should be set aside and mapped to custom codes in the format of the OMOP CDM concept table. During the ETL, implementers should have a test OMOP CDM database and fill it with a month's data from the source database. It is also important to identify the transformations that the source data needs to undergo during the ETL, such as an unpivot.

Lastly, we recommend using the Docker containers for the validation step if they are available. An alternative solution is to prepare a list of dependencies with their version before trying to run the application. Throughout the entire process, data security should be closely monitored. Test data should be carefully generated and anonymized. It should also be clear what access should be given for each application since some need more than others. If other health organizations, especially ICU quality registries, want to join the NICE in joining the European science community, we believe that our experiences can help with their transition.

References

[1] van der Lei J. Use and abuse of computer-stored medical records. Methods of information in medicine. 1991;30(02):79-80.

[2] OHDSI – Observational Health Data Sciences and Informatics [Internet]. Ohdsi.org. 2022. Available from: https://www.ohdsi.org/

[3] de Groot R, Cornet R, de Keizer N, Benis N, Raiez F. OMOP CDM compared to ContSys (ISO13940) to make data FAIR – OHDSI [Internet]. Ohdsi.org. 2022. Available from: https://www.ohdsi.org/2020-global-symposium-showcase-52/

[4] Dongelmans D, Pilcher D, Beane A, Soares M, del Pilar Arias Lopez M, Fernandez A et al. Linking of global intensive care (LOGIC): An international benchmarking in critical care initiative. Journal of Critical Care. 2020;60:305-310.

[5] NICE Data Dictionary [Internet]. NICE. 2022. Available from: https://www.stichting-nice.nl/dd/#start

[6] Informatics O. The Book of OHDSI [Internet]. Ohdsi.github.io. 2022. Available from: https://ohdsi.github.io/TheBookOfOhdsi/

[7] EHDEN Academy [Internet]. Academy.ehden.eu. 2022. Available from: https://academy.ehden.eu/

[8] OHDSI Common Data Model Wiki [Internet]. Ohdsi.github.io. 2022. Available from: https://ohdsi.github.io/CommonDataModel/index.html

[9] Observational Health Data Sciences and Informatics GitHub [Internet]. GitHub. 2022. Available from: https://github.com/OHDSI

[10] Zimmerman JE, Kramer AA, McNair DS, Malila FM. Acute Physiology and Chronic Health Evaluation (APACHE) IV: hospital mortality assessment for today's critically ill patients. Critical care medicine. 2006 May 1;34(5):1297-310.

[11] Philofsky M, RN, MS Mapping Custom Source Codes to Standard Concepts: A Comparison of Two Approaches [Poster]. 2020 Available from: https://www.ohdsi.org/2020-global-symposium-showcase-18/

[12] Tystan, Knoll C, Brandt P. Cohort Generation in Atlas with Generation status Failed [Internet]. OHDSI Forums. 2020. Available from: https://forums.ohdsi.org/t/cohort-generation-in-atlas- with-generation-status-failed/5690/31

[13] Evans L, Suchard M. OHDSI Broadsea, GitHub Repository [Internet]. 2021. Available from: https://github.com/OHDSI/Broadsea

Challenges of Trustable AI and Added-Value on Health
B. Séroussi et al. (Eds.)

doi:10.3233/SHTI220477

An Ontology and Data Converter from RDF to the i2b2 Data Model

Jules FASQUELLE-LOPEZ [a], Jean Louis RAISARO [a]
[a] *Lausanne University Hospital*

Abstract. In a national effort aiming at cross-hospitals data interoperability, the Swiss Personalized Health Network elected RDF as preferred data and meta-data representation format. Yet, most clinical research software solutions are not designed to interact with RDF databases. We present a modular Python toolkit allowing easy conversion from RDF graphs to i2b2, adaptable to other common data models (CDM) with reasonable efforts. The tool was designed with feedback from clinicians in both oncology and laboratory research.

Keywords. RDF, i2b2, SPHN, interoperability, semantic web

1. Introduction

The Swiss Personalized Health Network (SPHN) defined its three-pillar strategy for medical data interoperability [1]. As the first and second pillars aim at, respectively, developing a common semantic framework and formally representing meta-data and data using the Resource Description Framework (RDF)[2], the third pillar encompasses metadata and data transformation in a use-case dependent manner for cross-institutional projects. The present work provides a first prototype for a modular and versatile implementation of the third pillar objective.

We propose a RDF converter that can extract both metadata and data samples from RDF graphs and transform them into the destination common data model (CDM) while minimizing information loss. As a first instantiation of this framework, we focus on the i2b2 common data model [3]. i2b2 is currently in use in several Swiss university hospitals and can count on a large user base both in the US as well as in Europe. Prior work on this subject draw equivalences between RDF and i2b2 [4], or detail a graph pruning algorithm [5] which inspired one of our modules.

Created by Murphy et al. [3], Informatics for Integrating Biology and the Bedside (i2b2) offers a modular database system for storing clinical data divided in service cells. The ontology cell (ONT) compiles all ontology information as unique codes and relies on file-system-like paths to encapsulate concept hierarchies. Another cell, named data repository cell (CRC), holds the patient observations as instances of elements referenced in the ontology cell. Patient, data provider, encounter and other details are stored using additional dimension tables available in the i2b2 "star schema". Our converter outputs CSV files ready to push into the i2b2 database.

2. Methods

By definition, the main challenge in building any type of connector is the diversity of scenarios and how generic the tool should be. This work was initiated in 2020 as an ad-hoc development task for the MedCo[6] project. Indeed, being funded by the SPHN consortium, the MedCo pilot deployments make use of the nominal SPHN semantics and ontologies available as RDF knowledge graphs. At the same time, MedCo provides a network privacy-preserving analytics layer on top of distributed i2b2 databases.

We aimed at developing a tool that would be robust to both minor architecture changes in the RDF graphs and representation in the output format. This made a priority of getting the framework easily configurable and expandable. To achieve this, we designed a modular architecture where much of the CDM-related dependencies are pushed to external configurable profiles as much as possible. The Converter consists of two main modules: an **Ontology converter**, and a **Data samples converter**. While complementary, their use cases might differ in practice: collaborative studies typically feature a common ontology and local data samples.

The Ontology converter is itself designed as two components interfacing on a very simple data structure, as explained in the next subsections.

2.1. Ontology converter

The design work for ontology conversion boils down to extracting meaningful spanning trees from the RDF knowledge graph. Main concepts are discriminated and labeled as entry points from which graph explorations are performed.

Authors of [4] note that the *Subject - Predicate - Object* RDF base triple is similar to the i2b2 *Subject - Predicate - Object* principle. Yet, while with RDF an Object can itself be a Subject, i2b2 integrates a strict hierarchy between Concepts and Modifier values. In i2b2, Modifiers can only be subordinates to Concepts. This strict hierarchy is illustrated in Figure 1.

The RDF classes do not fit naturally into the two i2b2 categories (Concept, Modifier). We decided to consider as i2b2 Concepts only the medical concepts which could be instantiated without the need of an other concept as context. For example, a *Drug* observation for a patient does not bear a natural meaning outside of its context, say *Drug Allergy* or *Drug Prescription*. More formally, RDF classes that appear as a property of another class are a priori discarded from the list of entry points.

Yet, this list of classes is entirely configurable. It defines the entry points for the graph discovery.

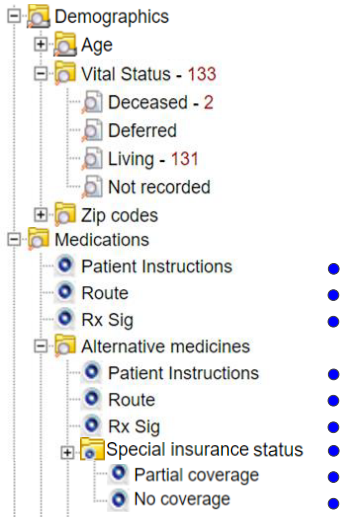
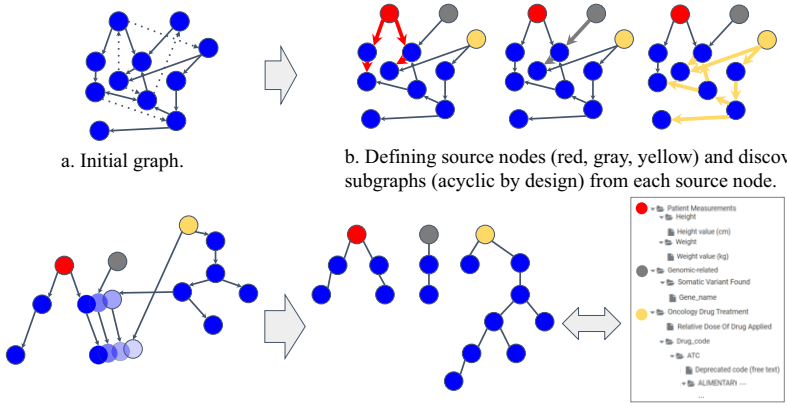


Figure 1. An example of i2b2 ontology where "modifiers" are indicated by a blue dot, every other element is a "concept". Modifiers are not mandatory, and when present are directly or indirectly subordinated to a concept. They can also be subordinated to an other modifier.

Extracting from RDF: We define a first module to extract a tree from the RDF ontology graph, abstracting all RDF-specific terminology into a simple parent/children interface. The steps of the algorithm are shown in Figure 2.

Starting from the main classes defined as "entry points" to the graph, a recursive walk is performed through specific edges only, in the same fashion as described in [5]. While Stöhr et al. were using `skos:topConceptOf/skos:hasTopConcept` to navigate the graph, we use by default the RDF Schema norm `rdfs:subClassOf/rdfs:domain/rdfs:range`. Our configuration files allow swapping these Unique Resource Identifier (URIs) for equivalent ones (if using an other RDF norm). By design of [1], the subgraph defined by such links does not contain cycles.

We project every visited RDF element onto two simple Python classes: `Concept` and `Property`. Each Concept instance contains an array of all Properties that apply to it (directly drawn from the RDF `rdfs:domain` predicate). Likewise, each Property instance contains an



a. Initial graph.

b. Defining source nodes (red, gray, yellow) and discover subgraphs (acyclic by design) from each source node.

c. List every path from each source node. Subpaths can then be explored more than once if shared by distinct sequences.

d. The resulting forest is topologically the same data structure as an i2b2 ontology (here displayed in the MedCo client).

Figure 2. The goal of the RDF parsing module is to unfold a graph (a) into a collection of trees (d).

array of all Concepts it points to (directly drawn from the RDF `rdfs:range` predicate). This interface captures the useful information from the RDF graph without its complexity. At the same time, the two-classes system prefigures the i2b2 Concept/Modifier model.

Both our Concept and Property classes implement a `get_children()` interface, returning the elements of their respective array. They also implement information interfaces allowing to retrieve either comments, language-specific descriptions of the item, or the URI of the original RDF item.

Drawing i2b2 metadata models: The i2b2-specific module fills one or several i2b2 metadata tables given a collection of Python objects implementing the `get_children()` interface.

In i2b2, every ontology item is characterized by its *path*, which is the concatenation of the path of its parent and its own name – which can optionally be different from its display name. Based on a tree built as described in the previous section, it is enough for this i2b2-specific module to visit every element in the tree top-down (using recursive calls to `get_children()`), constantly adding new *paths* to the output data frame.

A set of i2b2-specific routines using the other available interfaces (to determine a display name, a display icon, tooltips or a contextual menus, etc.) are also triggered to fulfill the i2b2 database norms. We define a trivial policy for Concept/Modifier distribution: the children of a python Concept (see subsection 2.1) yield modifiers unless the said concept is flagged as abstract, and all the children of a modifier are modifiers. Extending the Converter to a new data model implies writing a new CDM-specific module using the same interface. No querying of the RDF graph is needed.

2.2. *Data samples converter to i2b2*

Data loading occurs more frequently than ontology creation. This Data converter is then less complex than the Ontology converter. It implements direct links between RDF instances and i2b2 data tables.

The i2b2 data model implements SQL joins within the data repository (CRC) cell [3] using a unique *basecode*. The Ontology converter and the Data samples converter should independently generate matching basecodes.

In our design, the unique i2b2 basecode is drawn directly from the URI of the underlying RDF class or property (accessible through the Python interfaces) and its parent basecode. It encapsulates a *RDF-like* path at a very low level. Thanks to this design, the "data graph" already contains all necessary information to construct the basecodes, without embedding the whole "ontology graph".

Primary types such as text and numerical values encountered in the RDF resources are then added in the `value` fields of the appropriate table. The mappings between RDF types and the i2b2 specific tables and columns are entirely configurable.

2.3. Configuration

We provide three configuration files in JSON format. One is for the RDF ontology reading, one for the CDM-specific metadata writing and one for the data converter. They allow link configuration, concept blacklisting, etc. They also include lookup tables between the data model columns and the RDF data types (primary types (`xsd:string`, `xsd:double`) or more complex items (patient information, encounter details, etc).

3. Results

The expected output of the Converter is a full i2b2 schema as collection of mandatory tables in CSV format that can be pushed to an i2b2 database using simple SQL commands. As input, the Converter can take several ontology graphs and merge them if necessary. Terminology-specific graphs (such as ICD-10, LOINC, SNOMED-CT, etc.) are typically loaded separately in memory. In this implementation, all such terminologies are provided and maintained in RDF representation by the SPHN Data Coordination center [10]. We successfully tested the software against both the SPHN core ontology [9] and project-specific ontologies from the Swiss Precision Oncology and the Swiss BioRef SPHN-funded projects[11], which all feature large terminologies such as SNOMED-CT and ICD-10, for a total number of parsed items going over hundreds of thousands. All ontologies displayed in current MedCo demos and deployments were generated by the Converter based on RDF graphs issued by SPHN.

The outcome of the converter was validated from end-users and clinical researchers of the above-mentioned SPHN projects and the SPHN Data Coordination Center. Their feedback on the i2b2-formatted ontology allowed to define clearly which irrelevant RDF properties should be dropped, if macro-concepts should be created to increase user experience, etc. For instance, thanks to end-users' feedback, we integrated into the RDF to i2b2 ontology conversion process the codes' identifiers within the display names of all terminology items. This means the RDF ontology item of class *atc:A01A* with property *rdfs:label STOMATOLOGICAL PREPARATIONS* maps to the i2b2 element displayed as *ATC:A01A - STOMATOLOGICAL PREPARATIONS*. This allows both to leverage the automatic lexicographical ordering in i2b2's Web client and for end-users to search items by their code rather than by their label. The Converter is open-source and hosted on Github at the following address [12].

4. Discussion

The ontology Converter performs two distinct explorations: it first queries the RDF graph to construct a tree exposing an interface (2.1), and then walks the tree to craft i2b2 metadata entries (2.1). We consider this an acceptable trade-off between performance and modularity. Other works either implement direct pipelines from SPARQL (the query language for RDF) to i2b2 [4], [5], or cap the complexity of the supported RDF graphs [5]. With our design, adding support to a new common data model only requires replacing the i2b2-specific module, without touching the RDF parsing module. Moreover, an ontology converter is typically not a critical application in terms of performance requirements, as opposed to a data ETL.

A useful enhancement of the Converter would be handling of distant RDF graphs. For now, all graphs are loaded in the local memory of the machine running the tool. It is then limited by the RAM capacity. We do not either provide devops tools to create the output CSV tables on a distant machine (say, an i2b2 server).

Adding support of synonyms would also be a benefit, since both i2b2 and RDF feature such notions. The i2b2 storage system avoid duplications, which we do not do for now.

Prior work has built bridges between i2b2 and other data models, either through query translation [7] or full data model conversion pipelines [8]. The complexity of these tasks is partly due to the i2b2 specific representations that our interface aims to isolate. This advocates for a generic tool such at this one, that can be augmented for multi-model support.

5. Conclusion

The strength of our tool resides in its modularity and low-cost extensibility to new data models, but also in the fine-grained configuration of specific mappings. Designing the Ontology converter featuring a neutral tree interface also increases the tool's resilience to changes in the RDF schema.

Either it is used with the attached data converter or standalone, the ontology converter holds a predictable logic to map arbitrary large RDF knowledge graphs to the i2b2 common data model. Indeed, several hospitals participating in the SPHN initiative already have an i2b2 ETL pipeline from their data warehouse. In such cases, it would make sense to simply provide the converted i2b2 ontology for mapping local variables to i2b2 observations.

References

[1] Gaudet-Blavignac C, Raisaro JL, Touré V, Österle S, Crameri K, Lovis C. A National, Semantic-Driven, Three-Pillar Strategy to Enable Health Data Secondary Usage Interoperability for Research Within the Swiss Personalized Health Network: Methodological Study. JMIR Med Inform. 2021 Jun 24;9(6):e27591.

[2] RDF Schema 1.1 [Internet]. [cited 2022 Jan 18]. Available from: https://www.w3.org/TR/rdf-schema/

[3] Murphy SN, Weber G, Mendis M, Gainer V, Chueh HC, Churchill S, et al. Serving the enterprise and beyond with informatics for integrating biology and the bedside (i2b2). Journal of the American Medical Informatics Association. 2010 Mar 1;17(2):124–30.

[4] Solbrig HR, Hong N, Murphy SN, Jiang G. Automated Population of an i2b2 Clinical Data Warehouse using FHIR. AMIA Annu Symp Proc. 2018 Dec 5;2018:979–88.

[5] Stöhr, Hr MR, Majeed RW, Günther A. Metadata Import from RDF to i2b2. German Medical Data Sciences: A Learning Healthcare System. 2018;40–4.

[6] Raisaro JL, Troncoso-Pastoriza JR, Misbach M, Sousa JS, Pradervand S, Missiaglia E, et al. MedCo: Enabling Secure and Privacy-Preserving Exploration of Distributed Clinical and Genomic Data. IEEE/ACM Trans Comput Biol Bioinform. 2019 Aug;16(4):1328–41.

[7] Majeed R, Fischer P, Günther A. Accessing OMOP Common Data Model Repositories with the i2b2 Webclient – Algorithm for Automatic Query Translation. In: Studies in health technology and informatics. 2021.

[8] Klann JG, Joss MAH, Embree K, Murphy SN. Data model harmonization for the All Of Us Research Program: Transforming i2b2 data into the OMOP common data model. PLOS ONE. 2019 Feb 19;14(2):e0212463.

[9] The SPHN RDF Schema [Internet]. [cited 2022 Mar 31] Available from: https://www.biomedit.ch/rdf/sphn-ontology/sphn

[10] Krauss P, Touré V, Gnodtke K, Crameri K, Österle S. DCC Terminology Service—An Automated CI/CD Pipeline for Converting Clinical and Biomedical Terminologies in Graph Format for the Swiss Personalized Health Network. Applied Sciences. 2021 Jan;11(23):11311.

[11] SPHN Driver projects [Internet]. [cited 2022 Mar 31]. Available from: https://sphn.ch/network/projects/driver-projects/

[12] RDF-i2b2 converter, on CHUV Data Science Lab's GitHub, https://github.com/CHUV-DS/RDF-i2b2-converter.

Challenges of Trustable AI and Added-Value on Health
B. Séroussi et al. (Eds.)

doi:10.3233/SHTI220478

Aggregations of Substance in Virtual Drug Models Based on ISO/CEN Standards for Identification of Medicinal Products (IDMP)

Robert VANDER STICHELE[a,b,1] and Dipak KALRA[b]
[a] *Ghent University, Belgium*
[b] *European Institute for Innovation through Health Data, Belgium*

Abstract. In this study representation of chemical substances in IDMP is reviewed, with an exploration of aggregation levels for substance used in the virtual drug data models of RxNorm, SNOMED-CT, ATC/INN, and the Belgian SAM database, for products with a single substance and combinations of substances. Active moiety and available solid states forms are explored for diclofenac, amoxicillin, carbamazepine, amlodipine, with regard to their representation in coding systems such as WHODrug, SMS, UNII, CAS, and SNOMED-CT. By counting the number of medicinal products in Belgium for amlodipine in each level of aggregation, concepts for grouper of substances and two levels of grouper of medicinal products are illustrated. Recommendations are made for the further development of IDMP and its link to international drug classifications.

Keywords. Pharmaceutical preparations, active ingredient, ontology, semantic interoperability, standardization, RxNorm, Anatomical Chemical Therapeutic Classification, SNOMED-CT, International Non-proprietary Name, Identification of medicinal Product (IDMP).

1. Introduction

Precise identification of active substances in pharmacotherapeutic drugs is not an easy task. Historically, the majority medicines contained chemical substances. But, over the years, more and more new products are biologicals and other more complex substances.[1] Medicinal drugs may be either substances as their free base or acid or in a solid state form e.g., salt or co-crystal. (. Drug developers take decisions when choosing a solid state form for several reasons, such as solubility or absorbability. but it is not always clear from the labeling what exactly was chosen. Regulatory authorities, when approving chemical drugs should precisely describe the substance which does not always occur when considered clinically irrelevant. Nevertheless, the distinction of the solid state form has implications for the expression of strength of the product. Therefore, the ISO/CEN standards for identification of medicinal products (IDMP) require precise

[1] Corresponding Author, Robert Vander Stichele, Unit of Medical Informatics, Department of Public Health, Ghent University, 3K3, Depintelaan, 185, 9000 Gent, Belgium; E-mail: robert.vanderstichele@ugent.be.

identification of the active moiety and its solid state [2]. There are many coding systems to represent active moieties and their solid state forms, such as the International Non-proprietary Nomenclature (INN) from the World Health Organisation, the Chemical Abstract Service (CAS) Registry; Pubchem; DrugBank; SNOMED-CT; WHODRUG; and the substance registries of the FDA and EMA. The EU, with the support of FDA in various aspects, is currently engaged in a major cleansing operation, with systematic revision of over 30.000 registered substances with therapeutic activities, starting 2020, with a scheduled endpoint in 2022. As the European Substance Management System (SMS) is part of the pharmacovigilance system, it contains information on terms used to describe active substances and excipients actually reported in Individual Case Study Reports (ICSR) in authorized medicinal products from European and other countries. This provides valuable input for the central registry about variations in naming of substances (spelling errors, synonyms) and existence of solid state forms, so that correct and comprehensive cleansing can be performed. The aim of this study is 1) to illustrate the complexity of representation of chemical substance with 4 substances (diclofenac, amoxicillin, carbamazepine, amlodipine), 2) to explore the aggregation levels for substance used in the virtual drug data models of RxNorm, SNOMED-CT, ATC/INN, and the Belgian SAM database; 3) to explore possible approaches for representation of combinations of substances in medicinal products.

2. Methodology

Medicinal products authorized in Belgium were collected for amlodipine, amoxicilline, carbamazepine, and diclofenac. Only medicinal products containing single substances were included. The active moiety was identified, and an overview of relevant modifiers, if any. Fully identified (chemical) substances were attributed the role of Precise Active Ingredients (PAI) [2], namely an active moiety without modifier or the full list of one or more active moieties with modifiers [3]. The coding numbers were collected from WHODrug, SMS, UNII, CAS and SNOMED-CT. For concepts involving substance in virtual models in the of INN, ATC, RxNorm, SNOMED-CT, WHODrug, and SAM, the level of substance and product aggregation was defined. Finally, the approach to combinations of substances was analysed for ATC, SNOMED-CT, and SAM.

3. Results

3.1. Identification of the attribute of Precise Active Ingredient (PAI) for 4 substances

Substances data from SMS (EMA) and G-SRS (FDA)[4] were used, to list the possible modifiers for the selected substances amlodipine, amoxicillin, carbamazepine, and

[2] This term PAI is taken from the SNOMED Medicinal Product Specification V4.0 (http://snomed.org/mpm) and defined as : The substance that provides the therapeutic effect of the medicinal product, described using the fullest and most specific description of the substance as it is used in the product(s) being represented. This may include various Solid state forms , such as salts, and/or solvates.

[3] Acknowledgements: we thank Ursula Tschorn, Leonora Grandia, Julie James, Annet Rozema, Inti van Eck, Malin Fladvad for their contribution to table 1, as part of the UNICOM Action Program www.unicom.org

[4] https://www.fda.gov/industry/fda-data-standards-advisory-board/fdas-global-substance-registration-system

diclofenac. A choice has still to be made for a global codification and numbering system of substances, and a mechanism to attribute the role of PAI, which is essential for the production of a global identifier, the Pharmaceutical Product Identifier (PhPID), as proposed in the IDMP standard.[5] In Table 1, we listed the active moieties and the substances with the attribute of PAI for 4 substances, with the different coding numbers from WHODRUG, SMS, CAS, SNOMED-CT[6], as collected by the UNICOM Pilot Product List group. These concepts describe physical realities, with specific characteristics such as solubility, weight, etc.

From the SAM database, the medicinal products authorized and currently available on the market were selected. For amlodipine, 28 medicinal products packs from 8 marketing authorization holders (MAH) were selected, for carbamazepine 4 from 1 MAH, and for diclofenac 50 from 10 MAHs. For diclofenac, 4 salts were identified (no medicinal product was identified as the free acid as active moiety. as PAI); for amoxicillin two solid state forms; for carbamazepine there was no solid state described in a product; for amlodipine, three salt forms were identified: besilate, maleate, mesilate (only the 2 formers are on the market), although a lot of salts and salt hydrate forms are known for the amlodipine base.

Table 1. List of Moieties and Precise Active Ingredients

	WHODrug	SMS	UNII	CAS	SNOMED-CT
Active Moiety					
diclofenac	00372301001	100000092272	144O8QL0L1	15307-86-5	7034005 | Diclofenac (substance) |
amoxicillin (anhydrous, explicitly)	00249601145	100000091596	9EM05410Q9	26787-78-0	785686003 | Amoxicillin anhydrous (substance) |
carbamazepine	00052501001	100000092127	33CM23913M	298-46-4	387222003 | Carbamazepine (substance) |
amlodipine	00972401001	100000085259	1J444QC288	88150-42-9	386864001 | Amlodipine (substance) |
(Modified) substances with the attribute of Precise Active Ingredient					
diclofenac (iononized)	00372301001	100000092798	144O8QL0L1	15307-86-5	7034005 | Diclofenac (substance) |
dlclofenac sodium	00372302001	100000092272	QTG126297Q	15307-79-6	62039007 | Diclofenac sodium (substance) |
dlclofenac potassium	00372304001	100000092368	L4D5UA6CB4	15307-81-0	108515008 | Diclofenac potassium (substance) |
diclofenac diethylamine (syn. diclofenac diethylammonium)	00372303001	100000091074	6TGQ35Z71K	78213-16-8	426714006 | Diclofenac diethylammonium (substance) |
diclofenac epolamine	00372307001	100000085789	X5F8EKL9ZG	119623-66-4	425650004 | Diclofenac epolamine (substance) |
amoxicillin sodium	00249603001	100000090113	544Y3D6MYH	34642-77-8	427483001 | Amoxicillin sodium (substance) |
amoxicillin trihydrate	00249602001	100000092629	804826J2HU	61336-70-7	96068000 | Amoxicillin trihydrate (substance) |
carbamazepine	00052501001	100000092127	33CM23913M	298-46-4	387222003 | Carbamazepine (substance) |
amlodipine besilate	00972402001	100000090079	864V2Q084H	111470-99-6	84976003 | Amlodipine besilate (substance) |
amlodipine mesilate	00972404001	100000089571	291Y33EZHA	246852-12-0	not present
amlodipine benzoate	00972410001	Not existing	XD75TQ8A2P	1239916-29-0	789067004 | Amlodipine benzoate (substance) |
amlodipine maleate	00972403001	100000089370	CQ27G2BZJM	88150-47-4	421048000 | Amlodipine maleate (substance) |

[5] Once the cleansing results/advice is implemented, the EU will offer the US the match between the UNII code and the SMSID (formerly known as EUTCT code) to be loaded into the US Substance database (GSRS). The SMSID will also be fed into the European SPOR (Substance, Product, Organisation, Reference) multilingual terminological system to support IDMP implementation.

[6] https://who-umc.org/whodrug/whodrug-global/ ; https://www.ema.europa.eu/en/human-regulatory/research-development/data-medicines-iso-idmp-standards/spor-master-data/substance-product-data-management-services. . https://www.cas.org/ . https://www.fda.gov/industry/structured-product-labeling-resources/uniis-preferred-substance-names-and-their-identified-synonyms

[7] Acknowledgements: we thank Ursula Tschorn, Leonora Grandia, Julie James, Annet Rozema, Inti van Eck for their contribution to table 1, as part of the UNICOM action Program www.unicom.org

3.2. Analysis of aggregation of precise active ingredients and products in virtual drug data models

To proceed to aggregation, a distinction must be made between a grouping of substances and a grouping of products. A grouping of substances is allocating a collection of all substances that share the same moiety. When the substance has no solid state form (e.g. Carbamazepine, this substance will be identified as PAI and its group will have only one member. When the substance has one or more solid state forms, the group will contain all possible salts, hydrates or salt-hydrate substances.
A grouping of products with the same substance is a collection of all medicinal products with the same substance. In a first level of aggregation, the low level group of products can collect all medicinal products with the same substance attributed as PAI, as foreseen in the IDMP standard for the Pharmaceutical Product Identifier Level 1 (coding system yet to be established within IDMP). This group concept is very close to the SNOMED-CT abstraction of Medicinal Product Precisely (MP precisely). In a second high level of aggregation for the group of products, all products with the same active moiety (or with the same group of substances) can be collated in a "virtual therapeutic moiety", as implemented in the Belgian SAM database and in the SNOMED-CT concept "Medicinal Product Only (MP only)[7].

It is important to note that the same term "amlodipine" is used to describe the active moiety, the allocator of amlodipine solid state forms, and the group of products by the same active moiety of substances, the latter two to be represented by a different coding system than for the active moiety. In Table 2, this concept of grouping illustrated by the number of Medicinal Products (packs) on the Belgian Market.

Table 2. Distribution over medicinal products in Belgium over combinations of substance, dose form, and combinations for amlodipine (ATC C08CA01)

				Belgium	
Virtual Medicinal Product Group (INN prescription)				**MPs**	**MMPs**
Amlodipine oral 5mg					
	Pharmaceutical Product (PHPID Level IV)				
	amlodipine besilate	capsule, hard	5 mg	1	2
	amlodipine besilate	tablet	5 mg	6	9
	amlodipine mesilate	tablet	5 mg	0	0
	amlodpine maleaat	tablet	5 mg	1	2
Amlodipine oral 10 g					
	amlodipine besilate	capsule, hard	10 mg	1	2
	amlodipine besilate	tablet	10 mg	7	9
	amlodipine mesilate	tablet	10 mg	0	0
	amlodipine maleate	tablet	10 mg	1	2
	amlodipine maleate	coated tablet	10 mg	1	2

INN International Non-Proprietary Name Nomenclature **PhPID** Pharmaceutical Product Ide
MP : Medicinal Product **MPP** Medicinal Product Pack **MAH** Marketing Authorization Holder

[7]One could argue that the INN nomenclature could act as a terminology for the allocator of the active substance, and in fact it is used in the ATC classification as such. However, this is not an explicit function of the INN nomenclature, and there are only internal, not very robust coding numbers available. In the US based RxNorm system of the National Library of Medicine, little attention is paid to the substance solid state forms in medicinal products. Only high level grouping of substances is used at the virtual level (Semantic Clinical Drug concept). The IDMP model does not provide sophisticated levels of higher aggregation of Medicinal Products by substance, except for PhPID-Level 1.

3.3. Analysis of approaches for combinations of medicinal products

There are basically 3 approaches to deal with combinations of substances in Medicinal Products. The approach of the ATC Classification is to have separate fifth level classes for products with a single substance and for products with combinations (sometimes specified and sometimes not or designated by drug class names). In the Belgian SAM, combinations of substances are entered in the database at the same level of the single products, are limited to 3 substances (labeled a+b+c plus in case of more than 3 active ingredients, and linked to a separate database of single constituting substances with the attribute of PAI. SNOMED-CT has a mixed approach with MP Only (all products containing amlodipine as a single or in a specified combination) as a subclass of MP (all products containing amlodipine). It is still unclear which approach will be adopted in IDMP for handling drug combinations.

4. Discussion and Conclusions

With this work a proposal is made for the consistent use of substance terms with the attribute of precise active ingredient(s), either the active substance or its solid state form to build the PhPIDs of medicinal products. Several coding systems exist to describe these physical realities, and a choice will have to be made between one of these coding systems, or a new global system created. In addition, a terminology will have to be created with a coding system for the concept of grouping of substance, preferably as an extension of the INN Nomenclature. For the low level aggregation of medicinal products based on PAI, IDMP will have to provide a coding system for PhPID-Level 1. For the higher level aggregation of medicinal products based on the grouping of substances, also a coding system will need to be determined.[3] This higher level of aggregation is needed to operationalize the concept of INN prescriptions, and to make a precise link with the ATC classification. This could enhance the possibilities for international Computerized Decision Support Systems and Drug Information Resources to link more easily to the different national medicinal drug dictionaries. Precise identification of the substance in chemical substances as either the moiety or the substance including its solid state form is crucial to determining the precise active ingredient, as foreseen in the IDMP standards.

References

[1] Peryea T, Southall N, Miller M, Katzel D, Anderson N, Neyra J, Stemann S, Nguyễn ĐT, Amugoda D, Newatia A, Ghazzaoui R, Johanson E, Diederik H, Callahan L, Switzer F. Global Substance Registration System: consistent scientific descriptions for substances related to health. Nucleic Acids Res. 2021 Jan 8;49(D1):D1179-D1185.

[2] ISO 11238:2018 Health informatics -- Identification of medicinal products -- Data elements and structures for the unique identification and exchange of regulated information on substances. Snomed-CT Medicinal Product Model Specification V4.0.

[3] Van Bever E, Wirtz VJ, Azermai M, De Loof G, Christiaens T, Nicolas L, Van Bortel L, Vander Stichele R. Operational rules for the implementation of INN prescribing. Int J Med Inform. 2014 Jan;83(1):47-56.

Challenges of Trustable AI and Added-Value on Health
B. Séroussi et al. (Eds.)

doi:10.3233/SHTI220479

Non-Fungible Tokens as a Mechanism for Representing Patient Consent

James CUNNINGHAM[a,1], Nigel DAVIES[b], Sarah DEVANEY[c], Søren HOLM[c], Mike HARDING[b], Victoria NEUMANN[b] and John AINSWORTH[a]
[a] *Division of Informatics, Imaging and Data Sciences, Health EResearch Centre, Manchester Academic Health Science Centre, University of Manchester, Manchester, United Kingdom*
[b] *Department of Computing and Communications, Lancaster University, Lancaster, United Kingdom*
[c] *CSEP, Department of Law, School of Social Sciences, The University of Manchester, Manchester, United Kingdom*

Abstract. In recent years we have seen the adoption of distributed ledger technology (DLT), originally the mechanism underpinning the operation of the Bitcoin crypto currency, across a wider range of technology sectors including healthcare. DLT allows for the design of informatics systems with the properties of immutability, security, and decentralization. One recent innovation in the space has been the specification and development of Non-Fungible Tokens (NFTs). NFTs are decentralized DLT-based records that represent ownership of a unique digital asset. The predominant current use case for NFTs has been in the representation and sale of digital artwork, however the features offered by NFTs, unique-ness, immutability, transferability, and verifiability, are directly applicable to the design of health informatics systems. In this paper we explore these properties and describe a reference architecture for using NFTs as a means of representing and transferring records of patient's consent for medical data use.

Keywords. blockchain, health information technology, distributed ledger technologies, digital health

1. Introduction

Distributed Ledger Technology (DLT), or blockchain, is the core innovation of the Bitcoin digital currency, created by the pseudonymous Satoshi Nakamoto as a decentralized, trustless, and pseudonymous alternative to traditional fiat currencies [1]. DLT is the mechanism by which distributed nodes in a peer-to-peer network can achieve consensus as to a canonical sequence of transactions that have been processed by the network, even where there exist nodes within that network acting maliciously – a solution to the so-called Byzantine Generals problem [2]. Based around public key cryptography [3] DLT allows for the creation of digital networks with the properties of immutability, security, decentralization, and trustless-ness [4]. This technology has seen adoption across a range of industries including the energy sector [5], smart cities [6], government [7] and transport [8]. In healthcare DLT has seen use in a wide range of healthcare

[1] Corresponding Author, James Cunningham; E-mail: James.A.Cunningham@manchester.ac.uk.

applications such as vaccination coordination [9], medical record access [10], access to healthcare service marketplaces [11] and pharmaceutical supply chain management [12]. In all such instances, the key properties of DLT around security, data permanence and lack of central control, are being leveraged to enhance both institutional and public trust in the systems being developed and to provide guarantees around the security and integrity of sensitive medical data [11]. The core features of DLT have been extended with the adoption and implementation of smart contracts [13], executable specifications of instructions that can manipulate data held on a blockchain. The predominant use case of smart contracts has been the creation of digital tokens – fungible, ownable and tradable as-sets represented by a record within a smart contract of the quantity of the token owned network participant. These tokens are utilized for a range of purposes such as, for example, being sold for funding in exchange for an ownership stake in an enterprise, the representation of weighted voting rights in some form of governance mechanism or simply as currencies in and of themselves [14]. More recently we have seen the use and adoption of Non-Fungible Tokens (NFTs). Contrasted to standard tokens NFTs are used to represent a unique, rather than homogenous and fungible, asset, for example an item of digital artwork, whilst sharing the properties of ownership and tradability with ERC20 tokens [15]. The predominant use of NFTs has in-deed been within digital artwork, where the creators of artworks can embed and transfer ownership rights to digital art works/ This ability to imbue uniqueness onto inherently reproducible digital resources can be seen as the key innovation of NFTs. Other examples of NFTs include applications within the internet of things, to represent assets within land registries and ownership digital assets within online gaming. In healthcare specifically NFTs have been utilized as a potential governance mechanism for the governance of genomic data sets [15]. This paper explores the use of NFTs in the healthcare space, in particular the representation of consent for data re-use.

2. Methods

NFTs are implemented as coded specifications of smart contracts hosted on blockchain infrastructure. The pre-dominant implementation of NFTs is on the Ethereum blockchain, where NFTs are implemented against the ERC-721 standard. This defines NFTs as being smart contracts that record the following information for a given asset that is to be represented by the NFT: A record of the owner of the NFT; An optional quantity representing the quantity of the asset held by the corresponding owner; Functions for transferring the ownership of the NFT, and for delegating the ability to transfer ownership to third parties; A URI recording the location of the digital asset represented by the NFT.NFTs define a **unique** representation of a digital asset, that is, whilst the representation of the asset itself may be duplicated, there is only a single canonical representation of that asset as recorded by the underlying NFT. They are **verifiable**; given reference to an NFT, there is a defined method of deciding whether the representation being referenced is the same one as originally created by the originator of that NFT. They are **transferable**, meaning they are owned by the holder of a single private key and that **ownership** property can be transferred to the owner of a different private key. They exhibit **immutability**, in that information recorded about an NFT, including its current state and the history of its ownership cannot be deleted or changed in any way. These properties have broad applicability across applications in the health informatics domain. **Uniqueness** – references to medical data, and in particular records

of the associated rights to use that data should be singular items which cannot and should not be copied within the bounds of a system. The property of unique-ness allows such items to be represented without reliance on a centralized, potentially fallible data bases. **Verifiability** – it is crucial that the provenance and legitimacy of digital artifacts within healthcare systems is accounted for, Cryptographic verifiability can ensure that this property is met. **Transferability** – ownership of digital medical records, the rights to access those records and metadata about the use of those records can all be individually owned and transferable instances within healthcare applications. Immutability – the ability to audit medical systems to ensure regulatory compliance and to allow patients to verify legitimate use of their data is automatically accounted for by the property of immutability.

3. Results

We produced an architecture and reference implementation of a system for transmitting the consent of multiple data subjects to use their data for research purposes and to allow legitimate consumers of such data to apply these consents to obtain data from medical data providers. Once these consents have been obtained and used for a specified purpose then they are no longer able to be used to obtain further data. This system uses NFTs as the mechanism for recording and transmitting these bundles of consents between data consumers and data providers in such a way that the system does not need to rely on a trusted third party to verify the legitimacy of these consents. The system is designed to enable data subjects to record signed records of consent that permit Data Consumers to request medical data from Data Providers in line with the consents that have been given by the subjects to which that data pertains. Further: Consent should be obtainable independently of any entity other than the data subject who pro-vided that consent; Once obtained, consent can be used to legitimately obtain data by presenting it to a data provider which acts as the guardian of a data subject's medical data; There should be a permanently auditable record of which consents have been used by who for what purposes; No entity should be able to prevent or revoke the application of a consent to obtain data out-side of the data subject and the data provider. The high-level architecture of the system is illustrated in Figure 1, a description of the individual components of the system is given in the following sections.

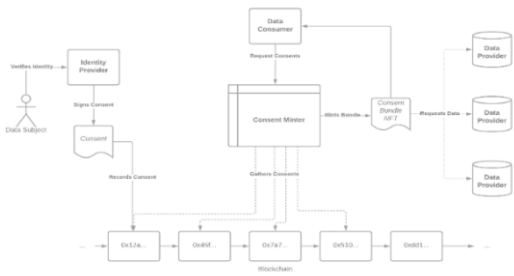


Figure 1. System Architecture.

Data Subject: an individual who has the rights to consent to the use of medical data for a given purpose. Within the system the role of a data subject is to verify their identity with an Identity Provider, allowing them access to the functionality of the system.

Identity Provider: the component responsible for verifying the identity of data subjects. An Identity Provider transmits and records signed records of Consent from Data Subjects.
Consent: an object consisting of a message, signed by an Identity Provider, which is written to the underlying blockchain through a smart contract, recording the following information: The signature of the Identity Provider issuing the consent; A pseudonymized identifier of the subject of the consent; A designation of the type of data the consent refers to; A designation of the purpose of use which is being consented for.
Data Consumer: entities within the system that wish to request data from Data Providers. Data Consumers can request data about sub-jects by providing references to recorded consents for the use of data to Data Providers in the form of Consent Bundles.
Consent Minter: Once consents have been gathered from data subjects and recorded, Data Consumers can call a Consent Minter smart contract which mints a unique NFT that references a series of recorded consents that have been issued for a given purpose.
Consent Bundle NFT: A Consent Bundle is an NFT representing a group of consents for a given purpose. The bundle can be passed directly from a Data Consumer to a Data Provider alongside a request for data to demonstrate the legitimacy of that request. Alternatively Consent Bundles can be passed or exchanged between data consumers or reserved for use at a later data. Once used to obtain data the NFT can be marked by the Data Provider as having been used. This can allow other data providers to base their decision to release data against that NFT on its previous use or can be 'burned' by the Data Provider, preventing its future use.
Data Provider: Data Providers are the entities within the system that are responsible for the storage of medical data and the transmission of that data to Data Consumers. They are responsible for verifying the identity of the Data Con-sumer making the request for data to ensure the legiti-macy of that request, and for consuming and processing the Consent Bundle object used to verify the legitimacy of an incoming request. The system was implemented on top of the Ethereum network. Since deployments and interactions with smart contracts on the network incur a so-called gas fee, pay-able in the Ethereum cryptocurrency, we deployed the system onto one of the Ethereum test networks, where obtaining the 'currency' used for gas payments is free. Smart contracts were written using version 0.8.0 of the Solidity programming language and the non-blockchain specific portions of the system were written in a mixture of the Go programming language and JavaScript, particularly utilizing the Web3.0 Ethereum libraries.

4. Discussion

We have looked at a reference implementation of a system whereby consented bundles of rights to access healthcare data are represented as NFTs and can be passed between parties without the knowledge of either the issuer or consumer of those rights. When those rights are exercised to access data those rights then expire. Whilst such a system could be implemented using a traditional informatics architecture, the ability to transfer privately, securely, and freely such rights around before access, and to do so outside of the knowledge and control any centralized entity are features unique to an DLT-based NFT solution. Whether these properties are necessary or even desirable in the context of an eHealth application will be dependent on the specific environment in which the system is implemented and will need to account for ethico-legal considerations. However, as with all healthcare systems, public perception and trust can be crucial factors in determining the ultimate success of a given application. The ability to transfer trust away

from a single entity into a system wide implementation may enhance both the perception of trust in the system and serve to mitigate the fallout from misuse or failure of the system away from a single entity thereby encouraging adoption. NFTs can provide a mechanism by which implementations of health informatics architectures can remove centralized sources of trust where it is deemed necessary. Where this is seen as beneficial and a desirable property NFTs provide an efficient and so-far unique means of doing so. NFTs represent a means of imparting uniqueness on digital assets and allow such assets to be owned and transferred as singular items (contrasted even with other blockchain-based approaches to managing patient consent), whilst preventing their duplication. To allow this in an informatics infrastructure that does not rely on the single entity or trusted third party to provide such functionality is a function unique to NFTs. Within medical informatics applications the properties provided by NFTs can potentially be used in situations that require the exchange and execution of unique one-time rights in situations where trust in the overall system will be enhanced by removal of reliance on a single component of the system. Moving forward we would conclude that NFTs can play a significant role in the development of eHealth applications.

References

[1] Nakamoto S. "A peer-to-peer electronic cash system." Bitcoin.–URL: https://bitcoin. org/bitcoin. pdf 4 (2008).
[2] Lamport L, Shostak R, and Pease M. "The Byzantine generals problem." In Concurrency: the Works of Leslie Lamport, (2019) pp. 203-226.
[3] Salomaa A. Public-key cryptography. (2013)
[4] de Leon DC, Stalick AQ, Jillepalli AA, Haney MA, Sheldon FT. Blockchain: properties and misconceptions. Asia Pacific Journal of Innovation and Entrepreneurship. 2017 Dec 4.
[5] Mihaylov M, Jurado S, Avellana N, Van Moffaert K, de Abril IM, Nowé A. NRGcoin: Virtual currency for trading of renewable energy in smart grids. In11th International conference on the European energy market (EEM14) 2014 May 28 (pp. 1-6). IEEE.
[6] Biswas K, Muthukkumarasamy V. Securing smart cities using blockchain technology. In2016 IEEE 18th international conference on high performance computing and communications; IEEE 14th international conference on smart city; IEEE 2nd international conference on data science and systems (HPCC/SmartCity/DSS) 2016 Dec 12 (pp. 1392-1393). IEEE.
[7] Atzori M. Blockchain technology and decentralized governance: Is the state still necessary?. Available at SSRN 2709713. 2015 Dec 1.
[8] Yuan Y, Wang FY. Towards blockchain-based intelligent transportation systems. In2016 IEEE 19th international conference on intelligent transportation systems (ITSC) 2016 Nov 1 (pp. 2663-2668). IEEE.
[9] Williams-Grut O. Estonia is using the technology behind bitcoin to secure 1 million health records. Bus Insid. 2016.
[10] Azaria A, Ekblaw A, Vieira T, Lippman A. Medrec: Using blockchain for medical data access and permission management. In2016 2nd international conference on open and big data (OBD) 2016 Aug 22 (pp. 25-30). IEEE.
[11] Leeming G, Cunningham J, Ainsworth J. A ledger of me: personalizing healthcare using blockchain technology. Frontiers in medicine. 2019 Jul 24;6:171.
[12] Mattke J, Hund A, Maier C, Weitzel T. How an Enterprise Blockchain Application in the US Pharmaceuticals Supply Chain is Saving Lives. MIS Quarterly Executive. 2019 Dec 1;18(4).
[13] Wood G. Ethereum: A secure decentralised generalised transaction ledger. Ethereum project yellow paper. 2014 Apr;151(2014):1-32..
[14] Victor F, Lüders BK. Measuring ethereum-based erc20 token networks. InInternational Conference on Financial Cryptography and Data Security 2019 Feb 18 (pp. 113-129). Springer, Cham.
[15] Wang Q, Li R, Wang Q, Chen S. Non-fungible token (NFT): Overview, evaluation, opportunities and challenges. arXiv preprint arXiv:2105.07447. 2021 May 16.
[16] Uribe D, Waters G. Privacy laws, genomic data and non-fungible tokens. The Journal of The British Blockchain Association. 2020 May 30:13164.

Challenges of Trustable AI and Added-Value on Health
B. Séroussi et al. (Eds.)

doi:10.3233/SHTI220480

A Conceptual Framework for Representing Events Under Public Health Surveillance

Anya OKHMATOVSKAIA[a,1] Yannan SHEN[a], Iris GANSER[a,b], Nigel COLLIER[c], Nicholas B. KING[a], Zaiqiao MENG[c] and David L. BUCKERIDGE[a]
[a] *School of Population and Global Health, McGill University, Canada*
[b] *Bordeaux Population Health Research Center, Bordeaux University, France*
[c] *Department of Theoretical & Applied Linguistics, University of Cambridge, UK*

Abstract. Information integration across multiple event-based surveillance (EBS) systems has been shown to improve global disease surveillance in experimental settings. In practice, however, integration does not occur due to the lack of a common conceptual framework for encoding data within EBS systems. We aim to address this gap by proposing a candidate conceptual framework for representing events and related concepts in the domain of public health surveillance.

Keywords. Event-based surveillance, conceptual modelling, outbreak detection

1. Introduction

Event-based surveillance (EBS) is the process of monitoring for, and reporting on, events of potential public health importance using information sources not specifically designed for this purpose, such as online news articles, social media, drug sales, or absenteeism data [1]. Compared to indicator-based surveillance, which relies on confirmed cases of known diseases, EBS is characterized by potentially higher sensitivity and timeliness, but usually at the cost of specificity. EBS systems face unique challenges, including questionable source reliability, low signal-to-noise ratio, and biases introduced by uneven internet coverage and trends in media reporting. Due to differences in scope, orientation, and design, the performance varies significantly between EBS systems and research has shown the benefit of integrating information across multiple systems [2]. However, for such integration to occur routinely in practice, semantic interoperability is necessary, i.e., the representation of key concepts must be aligned among systems.

At the core of all EBS systems is the concept of an *'event'*. While this concept appears intuitively clear, its use is inconsistent across different systems, particularly in relation to the terms *'signal'* and *'outbreak'*. A common approach is to view an event as a stage in the alerting process, for example: "Once the signal has been triaged and verified, it becomes an event" [3]. This narrative is ontologically problematic and assumes an implementation-specific event definition, thus making it impossible to compare different systems, as their decision processes vary. Another common narrative suggests that there are two different kinds of events: 1) an occurrence of public health significance (e.g. disease outbreak), and 2) other events (e.g. an increase in drug sales or Google searches)

[1] Corresponding Author, McGill University Clinical and Health Informatics, 2001 McGill College Avenue, Suite 1200, Montréal, Québec H3A 1G1, Canada; E-mail: anna.okhmatovskaia@mcgill.ca.

that may indicate or signal the occurrence of public health significance: "The goal of EBS is to detect unusual events that might signal an outbreak." [4]. However, there is no explicit model to describe how those other events relate to a disease outbreak.

Reaching consensus regarding these high-level concepts should create a foundation for achieving semantic interoperability amongst systems, improving global EBS. In this paper, we propose a conceptual framework for representing events and related concepts in the context of EBS that can serve as a foundation for addressing these issues. In this initial modelling stage, we do not assume a specific formal encoding of our framework.

2. Related work

Formally defining and representing the dynamic aspects of the world, such as events and processes, has been a subject of multiple disciplines from philosophy to linguistics to artificial intelligence. Yet the definitions of event vary across different frameworks, and numerous, often incompatible, concept hierarchies have been proposed [5].

2.1. Upper-level ontologies and foundation models of events

Events are positioned at the top level of the concept hierarchy and are typically defined as things that happen in time (perdurants, occurrents), as opposed to things that maintain their existence (endurants, continuants). Events are included in all major upper-level ontologies, but alternative terms are often used to refer to them, e.g. 'eventive perdurant' [6], 'transition' [7], 'process' [8]. Events usually imply change over time and are distinguished from 'qualities' and 'states', with the exception of some models that consider static states as a type of event [9]. Besides defining events, existing event frameworks typically discuss event typology, their spatiotemporal relations, mereology, participation of non-event entities, and causal relationships.

Most frameworks agree that events have temporal boundaries, i.e. a distinct start and end time points. Allen & Fergusson [10] point out the arbitrary nature of these points, while in basic formal ontology (BFO) [8] the definition requires them to "correspond to real discontinuities". The relation of events to places is sometimes modelled as direct [11], or indirect, via participating agents or objects [9, 12, 13].

Events can be comprised of other events as proper parts. Guizzardi et al. [12] suggest the existence of atomic events that cannot be further subdivided, but this view is not supported by other models, which point out that atomicity of events can only be assumed at a chosen level of temporal granularity. Some authors emphasize the importance of an observer in identifying event boundaries and parthood. According to [10], events do not really exist, but are the way by which agents classify the observed patterns of change.

Kaneiwa et al. [9] categorize events into types according to participating agents and objects (e.g. natural event vs. artificial event) and distinguish several semantic functions of events (state change, spatial existence change, etc.). Rich Event Ontology (REO) defines many subclasses of events, including life event, cognitive event, motion, etc. [14].

2.2. Domain applications of event models

Brown et al. have pointed out that despite the proliferation of software ontologies, most have focused on developing rich object hierarchies, but not event hierarchies [14]. Indeed, upper-level ontologies, e.g., [6, 8], only define the general aspects of events at the most

abstract level, and few ontologies extend this concept. REO is one example of such an extension, providing a rich structure of events at varying levels of specificity across many domains, and relating events to their key objects and participants [14]. Other domain applications of formal event models include the Music Ontology [11], the Time Event Ontology for representing clinical events [15], and Adverse Event Ontology [16]. Among domain-specific event ontologies, BioCaster Event Ontology [17] is the most relevant to the work presented here: this ontology built with an natural language processing focus extends Descriptive Ontology for Linguistic and Cognitive Engineering (DOLCE) to provide a multilingual vocabulary for extracting infectious disease outbreak information from media reports. Here, we build on this work to propose a conceptual event model with the aim of improving interoperability in public health surveillance.

3. Conceptual Framework Description

In the proposed conceptual framework, we intend to tease apart, define, and clarify the relationships among several phenomena, which are often described inconsistently and sometimes confused in the context of EBS: a) public health (PH) events that the systems are trying to detect (e.g. E. Coli outbreak at a restaurant due to tainted meat), b) individual cases of illness or clusters of cases that are part of a PH event, c) events that are causally related to an underlying PH event (e.g. school absences, increase in the sales of cold remedies), and d) documents about events (e.g. news articles, social media posts).

3.1. Definitions

Our definition of an *event* closely matches a definition of a 'process' in BFO, and is similar to 'eventive perdurant' in DOLCE. An *event* is an occurrent that exists in time, composed of temporal parts, and is a maximally connected spatiotemporal whole with bona fide beginnings and endings, corresponding to real discontinuities

A *public health event* is usually defined as an occurrence that involves disease or death above expected levels for a given time and place. Thus, a public health event can be formalized as a subtype of event that involves a *causative agent/hazard*, one or more *affected populations* as participants, and comprises *sub-events* (parts) that involve cases of illness or death. Although qualifying incidence of disease or death as being "above expected level" is central to the definition of a PH event, we do not formally capture this aspect because of its contextual and subjective nature.

We adopt a view that partitioning of public health events into sub-events depends on the context and the observer (e.g., a reporting news source). A sub-event can be an individual case of illness or a cluster of cases. Thus, some public health events are observed as a sequence of many sub-events, and some – as just one (e.g., a case of anthrax).

In addition to PH events and their sub-events, other subtypes of event can be relevant to representing and reasoning about PH events, when they are causally related to a PH event or any of its parts and can serve as signals for EBS; these are '*associated events*.'

3.2. Properties and relations

At a minimum, PH events are characterized by time, location, hazard type and affected population. Events unfold over a *time interval* with distinct and ordered *start time* and *end time*, and within certain geographic *location*, which are often unobservable directly.

PH events can be related to other events via mereological (part-of), temporal, and causal relations. Temporal relations between events of any granularity level are modelled in our framework using Allen's interval algebra. Recognizing the challenges of formally modelling causality, we propose to use a simple 'has-effect' relation for this purpose.

Various subcategories of PH event introduce additional properties. For example, infectious disease outbreaks are associated with a *disease* and a *pathogen*, from which other properties relevant to an outbreak can be inferred, such as the *mode of transmission*.

3.3. Event categories

The World Health Organization classifies PH events by hazard type into the following categories: zoonotic, chemical, food safety, infectious, natural disaster, nutritional deficiency, medical product, radio-nuclear, and undetermined [18]. Infectious disease outbreaks are the main focus of EBS systems, comprising 76%-93% of all events [18]. Within the class of infectious disease outbreaks, two broad types of events are noteworthy from a practical perspective: 1) relatively small communicable *disease clusters*, localized in time and space, and 2) *epidemics* (e.g., COVID-19) that extend over long time intervals and large geographic regions and have complex dynamics (changing geography, multiple waves). The distinction between these two types is not ontologically strict, but reflects two different workflows in EBS.

While almost any kind of event can become causally related to a PH event or its subevents, based on our analysis of existing EBS systems we include in our framework the following categories of associated events as particularly relevant: environmental events, natural and artificial (e.g. increase in mosquito activity, chemical plant incident); health-related behaviors (e.g. drug purchases); health care system events (e.g. hospitalizations, lab tests); interventions (e.g. travel ban); and social events (e.g. panic, protests). Existing ontologies can be used to further classify events within each of these categories.

3.4. Event documentation

Following [13], we believe that it is important to represent documents (e.g. media reports) about public health events, their sub-events, and associated events explicitly. Documents are information objects that represent events, and as such, they have a set of properties, distinct from that of events. Documents typically contain only partial and, possibly, inaccurate or distorted information about an event, increasing the possibility of introducing false positives or negatives. Also, multiple documents can refer to the same event, complementing or contradicting each other. A wide-spread issue in EBS of conflating events and event reports, poses a barrier to obtaining accurate event information.

4. Discussion and Future Work

Formal event models, including some intended for use in public health surveillance, have been proposed in the past. However, these models have not been adopted widely, possibly because their development was motivated by issues within individual systems, such as improving the parsing of news reports. Building on this prior work, our framework is developed based on the analysis of conceptual similarities and discrepancies among several existing EBS systems, aiming at creating a harmonized vocabulary with a potential for broader adoption. The alignment with a widely used upper ontology, BFO, should

facilitate further incorporation of relevant domain knowledge from numerous BFO-mapped ontologies for a variety of applications.

Although the development of this framework has benefited from our interactions with colleagues who work with EBS, an important next step is to engage with EBS system stakeholders to obtain feedback on the proposed model regarding its utility and potential for adoption. After consulting with stakeholders and revising the framework based on their feedback, we intend to proceed to encoding this framework as a formal ontology. We will also elaborate use cases for application of the ontology within and across existing EBS systems. In the longer term, we can see potential value in extending a typology of associated events and building a rich causal model to link associated events to PH events to support reasoning for outbreak detection.

References

[1] O'Shea J. Digital disease detection: A systematic review of event-based internet biosurveillance systems. Int J Med Inform. 2017 May;101:15-22.

[2] Barboza P, Vaillant L, Mawudeku A, Nelson NP, Hartley DM, Madoff LC, Linge JP, Collier N, Brownstein JS, Yangarber R, Astagneau P; Early Alerting Reporting Project Of The Global Health Security Initiative. Evaluation of epidemic intelligence systems integrated in the early alerting and reporting project for the detection of A/H5N1 influenza events. PLoS One. 2013;8(3):e57252.

[3] Balajee SA, Salyer SJ, Greene-Cramer B, Sadek M, Mounts AW. The practice of event-based surveillance: concept and methods. Global Security: Health, Science and Policy. 2021 Jan 1;6(1):1-9.

[4] CDC Division of Global Health Protection (DGHP). Event-based surveillance. https://www.cdc.gov/globalhealth/healthprotection/gddopscenter/how.html; cited Jan 26, 2022.

[5] Galton AP. The ontology of states, processes, and events. 2012

[6] Gangemi A, Guarino N, Masolo C, Oltramari A, Schneider L. Sweetening Ontologies with DOLCE. In: Gómez-Pérez A, Benjamins VR, editors. Knowledge Engineering and Knowledge Management: Ontologies and the Semantic Web: 13th International Conference, EKAW 2002. Proceedings. Berlin, Heidelberg: Springer; 2002. p. 166--181.

[7] Pustejovsky J. The syntax of event structure. cognition. 1991 Dec 1;41(1-3):47-81.

[8] Smith B, Kumar A, Bittner T. Basic formal ontology for bioinformatics. 2005

[9] Kaneiwa K, Iwazume M, Fukuda K. An upper ontology for event classifications and relations. InAustralasian Joint Conference on Artificial Intelligence 2007 Dec 2 (pp. 394-403). Springer, Berlin, Heidelberg.

[10] Allen JF, Ferguson G. Actions and events in interval temporal logic. Journal of logic and computation. 1994 Oct;4(5):531-79.

[11] Raimond Y, Abdallah SA, Sandler MB, Giasson F. The Music Ontology. InISMIR 2007 Sep 23 (Vol. 2007, p. 8th).

[12] Guizzardi G, Wagner G, de Almeida Falbo R, Guizzardi RS, Almeida JP. Towards ontological foundations for the conceptual modeling of events. International Conference on Conceptual Modeling 2013 Nov 11 (pp. 327-341). Springer, Berlin, Heidelberg.

[13] Scherp A, Franz T, Saathoff C, Staab S. A core ontology on events for representing occurrences in the real world. Multimedia Tools and Applications. 2012 May;58(2):293-331.

[14] Brown SW, Bonial C, Obrst L, Palmer M. The rich event ontology. In Proceedings of the Events and Stories in the News Workshop 2017 Aug (pp. 87-97).

[15] Li F, Du J, He Y, Song HY, Madkour M, Rao G, Xiang Y, Luo Y, Chen HW, Liu S, Wang L, Liu H, Xu H, Tao C. Time event ontology (TEO): to support semantic representation and reasoning of complex temporal relations of clinical events. J Am Med Inform Assoc. 2020 Jul 1;27(7):1046-1056.

[16] Ceusters W, Capolupo M, De Moor G, Devlies J, Smith B. An evolutionary approach to realism-based adverse event representations. Methods of information in medicine. 2011;50(01):62-73.

[17] Kawazoe A, Chanlekha H, Shigematsu M, Collier N. Structuring an event ontology for disease outbreak detection. InBMC bioinformatics 2008 Apr (Vol. 9, No. 3, pp. 1-8). BioMed Central.

[18] World Health Organization. Acute public health events assessed by WHO Regional Offices for Africa, the Americas and Europe under the International Health Regulations (2005) 2016 Report.

Challenges of Trustable AI and Added-Value on Health
B. Séroussi et al. (Eds.)

doi:10.3233/SHTI220481

Data-Driven Modeling of Randomized Controlled Trial Outcomes

Zhehuan Chen[a], Yilu Fang[a], Hao Liu[a] and Chunhua Weng[a,1]
[a] *Department of Biomedical Informatics, Columbia University, New York, NY, USA*

Abstract. Anecdotally, 38.5% of clinical outcome descriptions in randomized controlled trial publications contain complex text. Existing terminologies are insufficient to standardize outcomes and their measures, temporal attributes, quantitative metrics, and other attributes. In this study, we analyzed the semantic patterns in the outcome text in a sample of COVID-19 trials and presented a data-driven method for modeling outcomes. We conclude that a data-driven knowledge representation can benefit natural language processing of outcome text from published clinical studies.

Keywords. outcome, randomized controlled trials, knowledge representation

1. Introduction

As the volume of medical evidence expands quickly, it is imperative to enable scalable machine comprehension of medical evidence and increase its accessibility for patients, clinicians, and researchers. The Participant, Intervention, Comparator, Outcome (PICO) framework is widely adopted for retrieving medical evidence [1]. In this framework, the outcome specifies anticipated measures, improvements, or affects [2], such as "*prolongation of remission," "longer survival,"* and "*blood glucose levels*." Efforts using natural language processing (NLP) to automate outcome extraction have been growing [3][4][5]. Efforts have also been made to standardize the representation for outcomes. For example, EBM-NLP has defined outcome categories such as Physical Health, which includes Pain, Adverse Effects, and Mortality; Mental/Behavioral Impact, which includes Mental health, Participant Behavior, and Satisfaction with Care; and Non-health Outcome, which includes Quality of Intervention, and Resource Use, and Withdrawals from Study [6]. Zarin *et al.* classified outcome measures as domains, specific measurements, specific metrics used to characterize each participant's result, and methods for data aggregating for each outcome measure [7]. The taxonomy of patient outcome proposed by Wilson and Cleary is more focused on the outcomes related to or affecting health-related quality of life, with emphasis on the causal relationship between different health concepts [8]. Lin *et al.* has proposed an ontology for treatment outcome in cancer and defined high-level classes as Assessment tools, domain, Measure, Relationship, and Value type [9].

[1] Corresponding author, Dr. Chunhua Weng, Department of Biomedical Informatics, Columbia University, 622 W 168 Street PH-20, New York, NY 10027, USA; E-mail: chunhua@columbia.edu.

Despite these existing efforts, outcome text in RCT publications is still not amenable for computing due to their complexities and the fact that the aforementioned ontologies are used primarily for manual knowledge engineering rather than by automated NLP pipelines. There is an unmet need to develop an outcome text knowledge representation that is data standards-based and interoperable with NLP systems. In this paper, we present an original method to develop a data-driven knowledge representation for clinical outcomes using example COVID-19 RCT abstracts obtained from PubMed.

2. Methods

2.1. Data

We retrieved the abstracts of 50 COVID-19 RCT abstracts randomly selected using indexed metadata from the MEDLINE database. Following the standard definition from the PICO framework [1], we manually annotated the PICO statements. After removing the duplicate, vague (e.g., "*significance*" or "*common type*"), and coarse outcomes (e.g., "*reaching the primary outcome*"), and correcting grammatical errors and misclassified outcomes, we obtained 408 distinct outcome text snippets for complexity analysis.

2.2. Complexity Analysis

The complexity analysis includes syntactic and semantic complexity.

2.2.1. Syntactic Complexity

Each outcome text snippet is decomposed into phrases, where ScispaCy with the pre-trained model "en-core-sci-lg" is used for part-of-speech tagging, and the Berkeley Neural Parser (https://parser.kitaev.io/) is used for constituency parsing. To achieve the best interpretability of a biomedical concept, for outcomes that can be decomposed into phrases, we only consider the most granular noun phrases (i.e., NP) and verb phrases (i.e., VP) in them, while for the rest, the entire outcomes will be retained as phrases with their phrasal categories. We categorize and analyze the outcomes based on the number of phrases they contain and their syntactic patterns.

2.2.2. Semantic Complexity

We identify the semantic elements of the outcomes and tally the proportion of their presence and align their categories with those in the previously published outcome analyses or ontologies. For instance, "Measuring object" is adopted from the disease-treatment ontology and refers to an affected object [10]; "Specific measurement" (referring to "Outcome measurement") and "Specific metric" are selected from the PICO ontology (Cochrane PICO Ontology); "Time frame" is adopted from the ClinicalTrials.gov (https://prsinfo.clinicaltrials.gov/results_definitions.html); and "Boolean connector" and "Exclusion connector" are adopted from the complexity analysis of eligibility criteria from Ross *et al.* [11]. We create "Defining connector" and "Statistical method" semantic elements. We further extend the "Exclusion connector" to "Negation cue," referring to all possible negation cues. The explanation of each semantic element is as follows:

a. **Boolean connector**: "*and,*" "*or,*" comma, etc. Besides, "*with*" can be counted as a Boolean connector if it could be changed to AND with no information loss (e.g., "*hospitalization with intensive care*" can be converted to "*hospitalization AND intensive care*"). Also, partially specific lists can be viewed as OR statements (e.g., "*hematologic indicators including c-reactive protein*" could be converted to "*hematologic indicators OR c-reactive protein*").
b. **Negation cue**: the negation cues such as "*not*" and "*other than.*"
c. **Measuring object**: the entity or event to be measured, such as "*c-reactive protein,*" "*anxiety,*" and "*hospitalization.*"
d. **Defining connector**: the adjective, verb or prepositional phrases describing the measuring object. For example, in the outcome "*viral replication in cells infected with sars-cov-2,*" "*viral replication in cells*" is the measuring object and "*infected with sars-cov-2*" is the defining connector. In the outcome "*number of patients turning negative,*" "*number of patents*" is the measuring object, and "*turning negative*" is the defining connector.
e. **Specific measurement**: the measuring methods or instruments which can be used independently, such as "*assessment scale*" and "*6-min walk test.*"
f. **Specific metric**: the specific data (e.g., "*concentration*") for the assessment of the extent to which the outcome has been achieved.
g. **Statistical method**: the statistical method to present the result.
h. **Time frame**: temporal descriptors (e.g., "*at 15 days*" in "*clinical status at 15 days*"), references to temporal events (e.g., "*during hospitalization*") or a combination of both (e.g., "*combined adverse reactions 7 days after injection*").

3. Results

3.1. Syntactic Analysis

According to Figure 1, 38.5% of outcomes contained more than one phrase. Within this subgroup of outcomes, 67.5% of the outcomes follow the syntactic pattern "NP + NP." There can be as many as 7 phrases in an outcome. For outcomes grouped by their number of phrases, the most frequently observed syntactic pattern corresponds to outcomes that contain multiple noun phrases. Some examples are provided in Table 1. According to Table 2, we identified that elements "Measuring object", "Specific metrics" and "Specific measurement" can each be outcomes independently, while the remaining 38.5% of the outcomes all contained more than one semantic element. The most common semantic element combination was the measuring object with specific metric. The specific metric specifies the data to collect for the outcome. It either came right after the measuring object, such as "*antibody level*" (measuring object: "*antibody,*" specific metric: "*level*") and "*lymphocyte count*" (measuring object: "*lymphocyte,*" specific metric: "*count*"), or before the measuring object with a preposition in between, such as "*concentrations of multiple inflammatory molecules*" (measuring object: "*multiple inflammatory molecules,*" specific metric: "*concentrations*"). The second and third common semantic element combinations were the measuring object with a time frame (e.g., "*7-day adverse reactions*") and with a defining connector (e.g., "*coronavirus nucleic acid from throat and nasal swab,*" where the measuring object is "*coronavirus nucleic acid*" and the defining connector is "*throat and nasal swab*"), respectively. Furthermore, 46 (11.3%) of the outcomes were with more than two semantic elements.

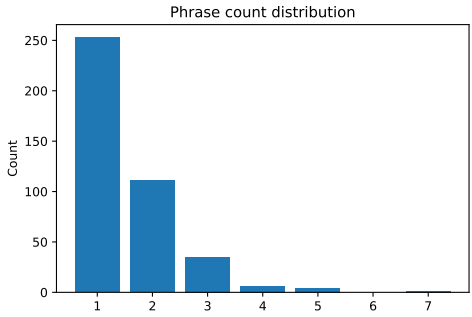


Figure 1. Phrase count distribution.

Table 1. Statistics of the syntactic patterns of the outcomes and their examples.

Phrase Count	Syntactic Patterns	Outcome Count (%)	Examples
1	NP	238 (58.3%)	*depression, symptom severity, t-cell counts*
	VP	10 (2.5%)	*subsequently died, discharged alive from hospital, cough, intubated*
	AdjP	3 (0.7%)	*not hospitalized, rash or itchy, severely ill*
2	NP + NP	106 (26.0%)	*clinical status at 11 days, duration of icu stay, ordinal scale of disease severity, time to fever resolution*
	NP + VP	5 (1.2%)	*time to discharge, case confirmed, patient withdrew*
3	NP + NP + NP	33 (8.1%)	*time to clinical improvement within 28 days, improvement from baseline of two points, change in symptom severity over 14 days*
	NP + NP + VP	2 (0.5%)	*number of patients turning negative, time from randomization to discharge*
4	NP + NP + NP + NP	6 (1.5%)	*time from starting the medication until discharge from hospital, percentages of patients with detectable viral rna at various time points*
5	NP + NP + NP + NP + NP	4 (1.0%)	*median number of days from symptom onset to start of study treatment*
7	NP + NP + NP + NP + NP + NP + NP	1 (0.2%)	*time from randomization to either an improvement of two points on a seven-category ordinal scale or discharge from the hospital*

3.2. Semantic Analysis

Table 2. Semantic patterns of the outcomes with one or two elements. For those outcomes with more than two semantic elements, they will be assigned to all possible semantic pattern groups with two elements.

Semantic Pattern	Outcome Count (%)
Measuring object alone	228 (55.9%)
Measuring object & Specific metric	130 (31.9%)
Measuring object & Time frame	46 (11.3%)
Measuring object & Defining connector	30 (7.4%)
Specific metric & Statistical method	24 (5.9%)
Specific measurement alone	19 (4.7%)
Measuring object & Boolean connector	15 (3.7%)
Measuring object & Negation cue	7 (1.7%)
Specific metric & Specific measurement	7 (1.7%)
Measuring object & Specific measurement	5 (1.2%)
Specific metric alone	4 (1.0%)

4. Discussion and Conclusion

A significant portion of outcomes contains more than one phrase or one semantic element, indicating the syntactic and semantic complexity in outcome text and implying the need to simplify outcome text. Approximately 38.5% of the outcomes contained more than one phrase, and almost the same portion of outcomes contained more than one semantic element, indicating phrase segmentation may potentially help reduce both the syntactic complexity and the semantic complexity of outcomes and improve the accuracy of their knowledge representations. For example, by phrase segmentation, the outcome "*respiratory secretion at day 4*" can be decomposed into two noun phrases "*respiratory secretion*" and "*at day 4*", where the former is semantically a measuring object and can be mapped to the standard observation concept "*Respiratory secretion*", and the latter is a time frame. This study has limitations. We only analyzed COVID-19 RCT abstracts. The generalizability of the semantic patterns to other disease domains remains to be tested. Second, the analyzed RCT abstracts are of a relatively small sample size. More evaluations are needed to test the completeness of knowledge in our representation. The PICO statements and the outcomes' semantic elements were manually annotated in this study; however, we have developed a parser to support the automatic recognition of the PICO statements [3]. A tool for semantic element extraction will be further developed.

References

[1] Richardson WS, Wilson MC, Nishikawa J, Hayward RS. The well-built clinical question: a key to evidence-based decisions. ACP J Club. 1995;123(3):A12-3.
[2] Davies KS. Formulating the evidence based practice question: a review of the frameworks. Evidence Based Library and Information Practice. 2011;6(2):75-80.
[3] Kang T, Zou S, Weng C. Pretraining to Recognize PICO Elements from Randomized Controlled Trial Literature. Stud Health Technol Inform. 2019;264:188-92.
[4] Kehl KL, Xu W, Lepisto E, Elmarakeby H, Hassett MJ, Van Allen EM, et al. Natural Language Processing to Ascertain Cancer Outcomes From Medical Oncologist Notes. JCO Clin Cancer Inform. 2020;4:680-90.
[5] Humbert‐Droz M, Izadi Z, Schmajuk G, Gianfrancesco M, Baker MC, Yazdany J, et al. Development of a natural language processing system for extracting rheumatoid arthritis outcomes from clinical notes using the national RISE registry. Arthritis Care & Research. 2022.
[6] Nye B, Li JJ, Patel R, Yang Y, Marshall IJ, Nenkova A, et al., editors. A corpus with multi-level annotations of patients, interventions and outcomes to support language processing for medical literature. Proceedings of the conference Association for Computational Linguistics Meeting; 2018: NIH Public Access.
[7] Zarin DA, Tse T, Williams RJ, Califf RM, Ide NC. The ClinicalTrials. gov results database—update and key issues. New England Journal of Medicine. 2011;364(9):852-60.
[8] Wilson IB, Cleary PD. Linking clinical variables with health-related quality of life: a conceptual model of patient outcomes. Jama. 1995;273(1):59-65.
[9] Lin FP, Groza T, Kocbek S, Antezana E, Epstein RJ. Cancer Care Treatment Outcome Ontology: A Novel Computable Ontology for Profiling Treatment Outcomes in Patients With Solid Tumors. JCO clinical cancer informatics. 2018;2:1-14.
[10] Khoo CS, Na J-C, Wang VW, Chan S. Developing an ontology for encoding disease treatment information in medical abstracts. DESIDOC Journal of Library & Information Technology. 2011;31(2).
[11] Ross J, Tu S, Carini S, Sim I. Analysis of eligibility criteria complexity in clinical trials. Summit on Translational Bioinformatics. 2010;2010:46.

Challenges of Trustable AI and Added-Value on Health
B. Séroussi et al. (Eds.)

doi:10.3233/SHTI220482

Thyroid Ultrasound-Image Dataset

Yasaman SHARIFI[a], Susan SHAFIEI[b], Morteza DanaiASHGZARI[c], Seyed Rasoul ZAKAVI[b] and Saeid ESLAMI[a,d,e,1]
[a]*Department of Medical Informatics, School of Medicine, Mashhad University of Medical Sciences, Mashhad, Iran*
[b]*Nuclear Medicine Research Center, Mashhad University of Medical Sciences, Mashhad, Iran*
[c]*Department of Computer Engineering, Faculty of Engineering, Islamic Azad University of Mashhad, Mashhad, Iran*
[d]*Pharmaceutical Research Center, Mashhad University of Medical Sciences, Mashhad, Iran*
[e]*Department of Medical Informatics, Amsterdam UMC (location AMC), University of Amsterdam, Amsterdam, Netherlands*

Abstract. Thyroid Computer-Aided Diagnosis (CAD) systems have been developed to assist radiologists in improving efficiency, reliability, and diagnosis performance. Often the performance of these CAD systems is evaluated with different datasets that make it incomparable. A valuable thyroid ultrasound (US) dataset is presented in this work. This dataset consists of 2450 thyroid US images from 2018 to 2020 in Prospective Epidemiological Research Studies in Mashhad, Iran (PERSIAN), a large national cohort study. These US images have the ROI of thyroid nodules and the associated American College of Radiology (ACR) Thyroid Imaging Reporting and Data System (TIRADS) features by expert physicians provided in XML format. Dataset's images are categorized into five groups based on the ACR-TIRADS (Tirads1-Tirads5). The presented dataset is expected to be a valuable resource to develop and assess thyroid CAD systems to help radiologists better diagnose.

Keywords. Thyroid, Sonography, TIRADS, Computer-Aided Diagnosis

1. Introduction

The prevalence of thyroid nodules is increasing worldwide, especially in women (67% of adults). But, the malignancy rate of these nodules is low (5-15%), and no invasive treatment such as fine-needle aspiration (FNA) or surgery will be required for all detected nodules [1-3]. Ultrasonography (US) is the best cost-effective and no invasive tool for thyroid nodules diagnostic and management. However, interpretation of thyroid US images has low reproducibility due to the noise and speckle of the ultrasound images and the different experience levels of physicians [4]. To overcome these challenges, Thyroid Imaging Reporting and Data System (TIRADS) proposes standard terminology for reporting thyroid nodules risk characteristics for effective management and avoiding unnecessary FNA (biopsy) [5]. Five TIRADS are internationally approved, which are ATA [6], ACE [7], EUTIRADS [8], KTIRADS [9], ACR-TIRADS [10]. Among these

[1] Corresponding Author, Saeid Eslami, School of Medicine, Vakil Abad Blvd, Mashhad, Iran; E-mail: EslamS@mums.ac.ir.

approved TIRADS, the American College of Radiology (ACR) has introduced ACR-TIRADS has the highest performance [11, 12]. However, the reproducibility of these systems was still about 45% -75% due to different experience levels of physicians and time-consuming evaluation of TIRADS factors. Still, interpreting US thyroid images and accurate scoring of thyroid nodules is challenging [4,13].

Artificial intelligence-based computer-aided diagnosis (CAD) systems can eliminate this operator dependency on interpreting thyroid US images and improve the diagnosis reliability and accuracy [14-18]. An actual limitation for developing such systems is the lack of thyroid US-image databases. Building this accessible publicly database plays an essential role in developing algorithms and supporting the comparison results and performances of different CAD systems.

This work presents a database of thyroid US images containing 2450 thyroid images with a complete annotation and associated diagnostic description based on the ACR-TIRADS lexicon performed by three experts. Accurate regions of interest (ROIs) square of nodules and their ACR-TIRADS properties provide in XML format.

2. Materials and Methods

2.1. Image acquisition

This dataset consists of 2450 ultrasound images from 1037 patients referred for thyroid ultrasonography from 2018 to 2020 in Prospective Epidemiological Research Studies in Mashhad, Iran (PERSIAN), a large national cohort study. This image was taken using a single system (Philips affinity 50G Ultrasound Machine-12.5 MHz linear transducer 5 cm) in sagittal and transverse plans and saved in a JPEG format. Images with at least one thyroid nodule were included in this dataset, and images with color Doppler flow and very low-quality images were excluded from the dataset.

Each thyroid nodule in the dataset was independently analyzed for ultrasonographic characteristics by three experienced physicians. Two independent physicians with 10 and 9 years of experience in thyroid US imaging reviewed images and reported associated ACR TIRADS features and levels of nodules (TR1–TR5) individually. To reach a consensus conclusion, they consulted with a third experienced physician with 18 years of experience in thyroid US imaging for any discrepancy in their result.

We developed a windows application for physicians to report and score the images based on ACR-TIRADS and draw square regions of interest (ROIs). This application was implemented using the C# language in the visual studio 2019 and SQL Server database. A snapshot of our application is shown in Figure 1. The menu on the left side of the screen includes an image list, and the selected image shows in the center of the screen. This application consists of five tabs, including:

1. Crop tab: in this tab, the physician can choose the appropriate image and draw a rectangle around the nodule present in each image (ROI).
2. Properties tab: in this tab, the physician can select the appropriate property of each ACR-TIRADS category from ComboBox, and by clicking on the "Calculate TIRADS" button, the TIRADS score and level of the nodule are calculated automatically, and all information is stored in the database, and the relevant XML file is created and saved.
3. Search tab: searches are based on each ACR-TIRADS attribute, and images can be stored in a specific folder.

4. Database tab: this tab displays all saved data in the database.
5. XML tab: in this tab, the user can create and save an XML file for each ACR-TIRADS category of images.

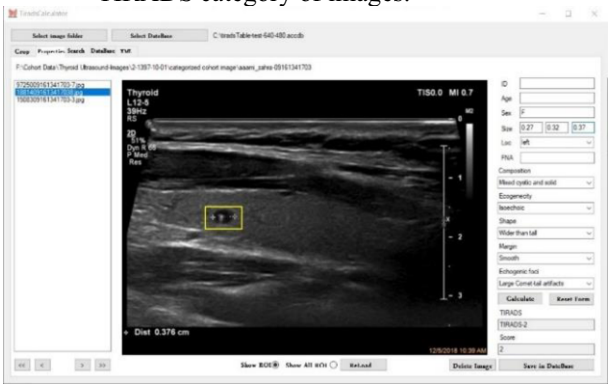


Figure 1. Tirads Calculator application.

2.2. Image Preprocessing

To make a relevant and valuable data set, image preprocessing is necessary. Image preprocessing of this work contains three steps, ROI extraction, artifact removal, and image resizing. Tirads Calculator performs ROI extraction of thyroid nodules. Since the original ultrasound images have patient information and a physician's caliper, we removed this inappropriate information and caliper by histogram analysis with python programming. We then resized the image size to 640×480 pixels.

3. Results

This dataset contains 2450 ultrasound thyroid images in JPEG format with relevant XML files. The size of the images was 640×480 pixels. The images are categorized into five groups based on ACR TIRADS (Tirads1-Tirads5).

Each XML file contains discriminant information such as the age and sex of patients, the size of the thyroid image, all the ACR-TIRADS properties, and coordinates of each nodule in the image.

Table 1. The number of images in each ACR-TIRADS level

Level	TIRADS-1	TIRADS-2	TIRADS-3	TIRADS-4	TIRADS-5
No	646	506	696	933	174

The age distribution of 1037 patients was 52.07±11.68 (mean±standard deviation) years, and the age range was between 29 and 89 years. Gender ratio was 60.17% (n=624) female and 39.82% (n=413) male. The number of images in each ACR-TIRADS level is shown in Table 1. Examples of original images and images without caliper and inappropriate information and sample images with ROI rectangles drawn by Tirads Calculator are represented in Figure 2.

4. Discussion

Interpretation of thyroid ultrasound images is a time-consuming task and has inter-observer variability due to differences in' physicians' experience and the nature of ultrasound images. Artificial intelligence-based CAD systems can eliminate this operator dependency on the interpretation of thyroid US images and improve the diagnosis reliability and accuracy. Still, the lack of thyroid US-image databases is a fundamental limitation for developing such systems. Lina Pedraza et al. (1) provides a dataset that contains 347 B mode ultrasound images from 299 patients (29 men, 270 women) scored by two observers in TIRADS. Their XML files contain classification data on two classes, benign and malignant. Our proposed paper presents 2450 Ultrasound images from 1037 patients (413 men, 624 Women) in jpeg format. This work produces a valuable thyroid US dataset with all associated ACR-TIRADS features and ROI of thyroid nodules by three expert physicians in the XML file. This dataset is ANN-ready with no artificial artifacts and is usable for all image analysis tasks in detecting and classifying thyroid nodules. It supports the performance comparison of the different CAD systems to help the expert improve the accurate diagnosis and decrease interpretation variation and unnecessary thyroid biopsies.

.

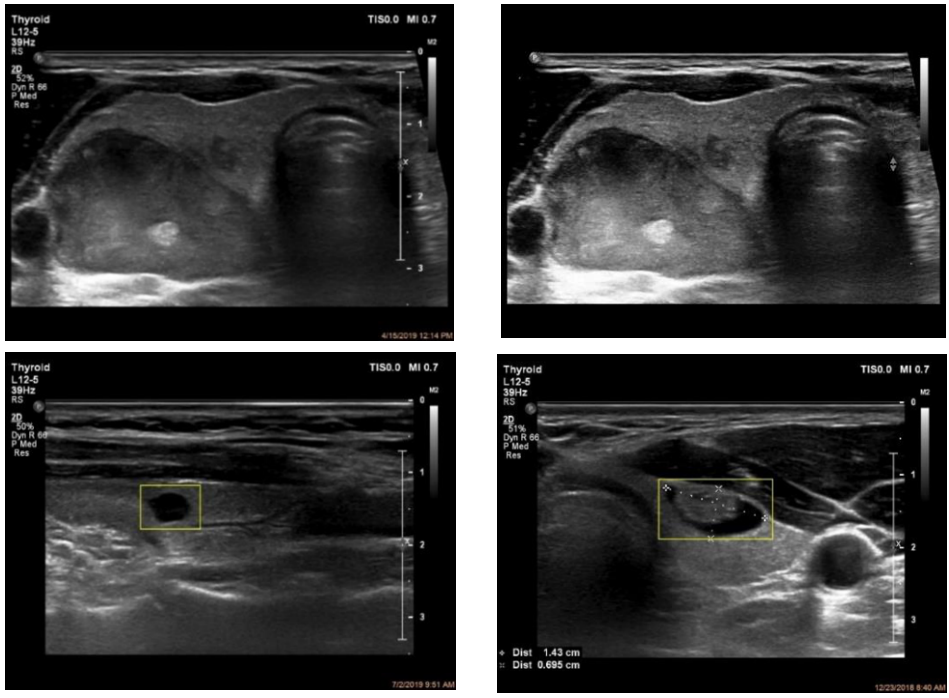


Figure 2. Example of original images and images after removing symptoms (above) and images with ROI rectangles drawn by Tirads Calculator (below)

5. Availability of Data and Materials

A sample of data described in this data note can be freely and openly accessed on the Harvard data server under (https://doi.org/10.7910/DVN/5E6IFH), and the whole database is accessible upon request.

References

[1] Khachnaoui H, Guetari R, Khlifa N, editors. A review on deep learning in thyroid ultrasound computer-assisted diagnosis systems. 2018 IEEE International Conference on Image Processing, Applications and Systems (IPAS); 2018: IEEE.

[2] Ren J, Baek JH, Chung SR, Choi YJ, Jung CK, Lee JH. Degenerating thyroid nodules: ultrasound diagnosis, clinical significance, and management. Korean journal of radiology. 2019;20(6):947-55.

[3] Zhuang Y, Li C, Hua Z, Chen K, Lin JL. A novel TIRADS of US classification. Biomedical engineering online. 2018;17(1):1-17.

[4] Grani G, Lamartina L, Cantisani V, Maranghi M, Lucia P, Durante C. Interobserver agreement of various thyroid imaging reporting and data systems. Endocrine connections. 2018;7(1):1-7.

[5] Horvath E, Majlis S, Rossi R, Franco C, Niedmann JP, Castro A, et al. An ultrasonogram reporting system for thyroid nodules stratifying cancer risk for clinical management. The Journal of Clinical Endocrinology & Metabolism. 2009;94(5):1748-51.

[6] Haugen BR, Alexander EK, Bible KC, Doherty GM, Mandel SJ, Nikiforov YE, et al. 2015 American Thyroid Association management guidelines for adult patients with thyroid nodules and differentiated thyroid cancer: the American Thyroid Association guidelines task force on thyroid nodules and differentiated thyroid cancer. Thyroid. 2016;26(1):1-133.

[7] Gharib H, Papini E, Garber JR, Duick DS, Harrell RM, Hegedus L, et al. American Association of Clinical Endocrinologists, American College of Endocrinology, and Associazione Medici Endocrinologi Medical Guidelines for Clinical Practice for the Diagnosis and Management of Thyroid Nodules-2016 Update Appendix. Endocrine practice. 2016; 22:1-60.

[8] Russ G, Bonnema SJ, Erdogan MF, Durante C, Ngu R, Leenhardt L. European Thyroid Association guidelines for ultrasound malignancy risk stratification of thyroid nodules in adults: the EU-TIRADS. European thyroid journal. 2017;6(5):225-37.

[9] Shin JH, Baek JH, Chung J, Ha EJ, Kim J-h, Lee YH, et al. Ultrasonography diagnosis and imaging-based management of thyroid nodules: revised Korean Society of Thyroid Radiology consensus statement and recommendations. Korean journal of radiology. 2016;17(3):370-95.

[10] Tessler FN, Middleton WD, Grant EG, Hoang JK, Berland LL, Teefey SA, et al. ACR thyroid imaging, reporting and data system (TI-RADS): white paper of the ACR TI-RADS committee. Journal of the American college of radiology. 2017;14(5):587-95.

[11] Griffin AS, Mitsky J, Rawal U, Bronner AJ, Tessler FN, Hoang JK. Improved quality of thyroid ultrasound reports after implementation of the ACR thyroid imaging reporting and data system nodule lexicon and risk stratification system. Journal of the American College of Radiology. 2018;15(5):743-8.

[12] Hoang JK, Middleton WD, Farjat AE, Langer JE, Reading CC, Teefey SA, et al. Reduction in thyroid nodule biopsies and improved accuracy with American College of Radiology Thyroid Imaging Reporting and Data System. Radiology. 2018;287(1):185-93.

[13] Sharifi Y, Shafiei S, Tabesh H, Aminzadeh B, Layegh P, Mahmoodzadeh A, Zakavi SR, Eslami S. Observation Variation in Ultrasonography Assessment of Thyroid Nodules. Asia Ocean J Nucl Med Biol. 2022; 10(1): 28-35.

[14] Chang Y, Paul AK, Kim N, Baek JH, Choi YJ, Ha EJ, et al. Computer‐aided diagnosis for classifying benign versus malignant thyroid nodules based on ultrasound images: A comparison with radiologist‐based assessments. Medical physics. 2016;43(1):554-67.

[15] Choi YJ, Baek JH, Park HS, Shim WH, Kim TY, Shong YK, et al. A computer-aided diagnosis system using artificial intelligence for the diagnosis and characterization of thyroid nodules on ultrasound: initial clinical assessment. Thyroid. 2017;27(4):546-52.

[16] Nugroho HA, Frannita EL, Ardiyanto I, Choridah L. Computer aided diagnosis for thyroid cancer system based on internal and external characteristics. Journal of King Saud University-Computer and Information Sciences. 2021;33(3):329-39.

[17] Prochazka A, Gulati S, Holinka S, Smutek D. Classification of thyroid nodules in ultrasound images using direction-independent features extracted by two-threshold binary decomposition. Technology in cancer research & treatment. 2019;18:1533033819830748.
[18] Sharifi Y, Eslami S, Danai Ashgzari M, Sargolzaei M. Deep Learning Applications in Analyzing Ultrasound Images of Thyroid Nodules: Protocol for Systematic Review. Front Health Inform. 2020; 9:33.
[19] Pedraza L, Vargas C, Narváez F, Durán O, Muñoz E, Romero E, editors. An open access thyroid ultrasound image database. 10th International Symposium on Medical Information Processing and Analysis; 2015: International Society for Optics and Photonics.

Challenges of Trustable AI and Added-Value on Health
B. Séroussi et al. (Eds.)

doi:10.3233/SHTI220483

WikiMeSH: Multi Lingual MeSH Translations via Wikipedia

Mikaël DUSENNE[a], Kévin BILLEY[a], Florent DESGRIPPES[c], Arriel BENIS[d], Stéfan Jacques DARMONI[a,b] and Julien GROSJEAN[a,b,1]
[a] *Department of Digital Health, Rouen University Hospital, France*
[b] *Laboratoire d'Informatique Médicale et d'Ingénierie des Connaissances en e-Santé (LIMICS), U1142, INSERM, Sorbonne Université, Paris, France*
[c] *LIFEN company, France*
[d] *Faculty of Industrial Engineering and Technology Management, Holon Institute of Technology, Holon, Israel*

Abstract. Objective: The aim of this paper is to propose an extended translation of the MeSH thesaurus based on Wikipedia pages. Methods: A mapping was realized between each MeSH descriptor (preferred terms and synonyms) and corresponding Wikipedia pages. Results: A tool called "WikiMeSH" has been developed. Among the top 20 languages of this study, seven have currently no MeSH translations: Arabic, Catalan, Farsi (Iran), Mandarin Chinese, Korean, Serbian, and Ukrainian. For these seven languages, WikiMeSH is proposing a translation for 47% for Arabic to 34% for Serbian. Conclusion: WikiMeSH is an interesting tool to translate the MeSH thesaurus and other health terminologies and ontologies based on a mapping to Wikipedia pages.

Keywords. MeSH, translating, Wikipedia

1. Introduction

The Medical Subject Headings (MeSH) thesaurus is a controlled and multi-hierarchically organized vocabulary produced by the National Library of Medicine (NLM). It is used for indexing, cataloging, and searching of biomedical and health-related information. MeSH includes the subject headings appearing in MEDLINE/PubMed [1]. Currently, the MeSH is available in 16 languages [2]. In its 2021 version, the MeSH thesaurus contains 29,754 Descriptors. Wikipedia [3] is a free content, multilingual online encyclopedia written and maintained by a community of volunteers through a model of open collaboration, using a wiki-based editing system. The domain Wikipedia.com was created in January 2001. Currently, it contains over 6,400,000 articles in English. All these articles are manually created and maintained. Wikipedia exists in 325 languages in the world. The aim of this paper is to propose an extended translation of the MeSH thesaurus based on Wikipedia pages using a new tool called "WikiMeSH". To our knowledge, no prior work has tested Wikipedia to enhance the multi lingual translation of a reference health terminology: in our example, the MeSH thesaurus. This work has

[1] Corresponding Author, Julien Grosjean; E-mail: julien.grosjean@chu-rouen.fr.

been partially granted by the European project (granted by Horizon 2020) HOspital SMART development based on Artificial Intelligence (AI; HosmartAI).

2. Methods

The goal of this WikiMeSH tool is to find Wikipedia entries in different languages for the descriptors of the MeSH thesaurus. The Wikipedia API allows to find the linguistic links (or mappings) for a given page, allowing for the exploration of a given topic in different languages. For each MeSH descriptor, a search for the corresponding entry on Wikipedia was performed.

3. Results

The WikiMeSH tool has been written in Python 3X and calls the Wikipedia API through the dedicated endpoints depending on the languages. Overall, the WikiMeSH tool was able to map a Wikipedia page for 18,191 MeSH descriptors (61.14%), which means that less than 39% of the MeSH descriptors have no Wikipedia links. 15,268 MeSH descriptors obtained a link based on their preferred term and 2,923 based on a synonym. The average number of detected Wikipedia languages per MeSH descriptor is 17.57 ± 34.80 (min-max: 0 - 300). The average number of added languages thanks to WikiMeSH is 17.61 ± 34.75. Among the top 20 languages of this study, seven are not already officially present in the MeSH thesaurus translations: Arabic, Catalan, Farsi (Iran), Mandarin Chinese, Korean, Serbian, and Ukrainian. For these seven languages, WikiMeSH is proposing a translation for 13,959 out of 29,754 MeSH descriptors (46.91%) for Arabic, 12,039 for Farsi (40.46%), 11,945 for Mandarin (40.16%), 10,839 for Ukrainian (36.43%), 10,770 for Catalan (36.20%), and 10,103 for Serbian (33.96%). The WikiMeSH results are integrated into the crosslingual terminology server HeTOP [4].

4. Conclusion

WikiMeSH is an interesting tool to translate the MeSH thesaurus and other health terminologies and ontologies based on a mapping to Wikipedia pages.

References

[1] Welcome to Medical Subject Headings. Available at https://www.nlm.nih.gov/mesh/meshhome.html (Accessed January, 10 2022).
[2] MSH (MeSH) – Synopsis. Available at: https://www.nlm.nih.gov/research/umls/sourcereleasedocs/current/MSH/index.html (Accessed January, 10 2022).
[3] Wikipedia. Available at: https://en.wikipedia.org/wiki/Wikipedia (Accessed January, 11 2022).
[4] Grosjean J, Merabti T, Dahamna B, Kergourlay I, Thirion B, Soualmia LF, Darmoni SJ. Health multi-terminology portal: a semantic added-value for patient safety. Stud Health Technol Inform. 2011;166:129-38. PMID: 21685618.

Challenges of Trustable AI and Added-Value on Health
B. Séroussi et al. (Eds.)

doi:10.3233/SHTI220484

Evaluation and Challenges of Medical Procedure Data Harmonization to SNOMED-CT for Observational Research

Ines REINECKE[a,1],Michael KALLFELZ[b], Martin SEDLMAYR[a], Joscha SIEBEL[a] and Franziska BATHELT[a]
[a] *Carl Gustav Carus Faculty of Medicine, Center for Medical Informatics, Institute for Medical Informatics and Biometry, Technische Universität Dresden, Dresden, Germany*
[b] *Odysseus Data Services GmbH, Berlin, Germany*

Abstract. The relevance of health data research on real world data (RWD) is increasing. To prepare national RWD for international research, harmonization with standard terminologies is required. In this paper, we evaluate to what extent the German OPS vocabulary in OHDSI covers codes present in RWD and mappings to SNOMED-CT. The evaluation identified a mapping gap of 21.1% in the RWD set.

Keywords. OHDSI, OMOP, SNOMED-CT, OPS, interoperability

1. Introduction

Real world data (RWD) empowers researchers to identify patterns in large healthcare data sets with the purpose to improve diagnoses and treatments. Hence, the relevance of observational research networks is increasing in Europe and around the globe [1-3]. The Observational Medical Outcomes Partnership (OMOP) CDM has become widespread in the recent years and can ensure the transferability and the comparability of results [4]. Mapping of national terminologies to international standards is crucial for participation in research networks on OMOP. This paper aims to evaluate the coverage of German OPS codes present in a RWD set to standardized vocabularies provided by OHDSI.

2. Methods

The procedure data was provided by the University hospital Dresden (UKD). The data set contains around 1.8 million procedure codes. Data analysis consisted of (a) calculating the mapping relationship between OPS procedures and SNOMED-CT concepts to identify mapping relationships and mappings other than mapping exactly one OPS to one SNOMED-CT code and (b) quantitative analysis to determine the frequency of procedure occurrence and the details on the RWD vocabulary coverage.

[1] Corresponding Author, Ines Reinecke, Carl Gustav Carus Faculty of Medicine, Center for Medical Informatics, Institute for Medical Informatics and Biometry, Technische Universität Dresden, Dresden, Germany; E-mail: ines.reinecke@tu-dresden.de

3. Results

The OPS vocabulary contains mappings to SNOMED-CT for 8,149 OPS codes (Table 1). The RWD under review contains 14,041 distinct OPS codes. The vocabulary does not provide a mapping for 10,044 OPS codes that have been present in the data set. Taking the frequency of the recorded procedures in the data set into account, the "no map" group represents 21.1% of the complete data set. 60.3% of the procedure data in the RWD set is covered by the OPS vocabulary with a 1-1 mapping, with a small number of concepts covering a large percentage of the data.

Table 1. Descriptive statistics of procedure usage and vocabulary details by mapping

mapping	OPS codes (vocabulary)	OPS codes (RWD)	procedure occurrences		descriptive statistics of procedure code frequency			
			total	percentage	mean	min	max	std
1-1	7493	3,413	1,113,835	60.3%	326	1	104,277	2,452
1-2	552	497	258,737	14.0%	521	1	26,082	1,740
1-3	92	77	83,207	4,5%	1081	1	24,780	3,403
1-4	12	10	1,820	0,1%	182	4	1,047	308
no map	30,306	10,044	390,496	21.1%	39	1	26,620	366
total	38,455	14,041	1,848,095	100%	132	1	10,4277	1,324

4. Discussion

We identified a substantial gap in the coverage of OPS code mapping to standard OHDSI concepts when compared to RWD from the UKD. Mapping data based on the currently available OPS vocabulary expose the risk of information loss. For the purpose of data harmonization, it is crucial for all countries to identify, name, understand and assess those risks when using data translated from national to international terminologies. Our next steps will include implementation of additional mappings based on the most frequent OPS procedure codes present in RWD to minimize the known gaps.

5. Conclusion

In this paper we have shown to what extent the RWD set is covered by the existing OPS vocabulary and its mappings to SNOMED-CT. Next steps have been proposed to proceed with closing identified gaps. This work is funded by the German Ministry of Education and Research (FKZ 01ZZ1801A/L).

References

[1] EMA DARWIN [Internet]. [cited 2022 Jan 17]. Available from: https://bit.ly/EMA-DARWIN.
[2] European Health Data Evidence Network – ehden.eu [Internet]. [cited 2022 Jan 17]. Available from: https://www.ehden.eu/.
[3] Schuemie MJ, et al. Principles of Large-scale Evidence Generation and Evaluation across a Network of Databases. J Am Med Inform Assoc. 2020;27:1331–1337.
[4] Reinecke I, Zoch M, Reich C, et al. The Usage of OHDSI OMOP - A Scoping Review. Stud Health Technol Inform. 2021;283:95–103.

Challenges of Trustable AI and Added-Value on Health
B. Séroussi et al. (Eds.)

doi:10.3233/SHTI220485

An Ontology for Cardiothoracic Surgical Education and Clinical Data Analytics

Maryam PANAHIAZAR[a,b,1], Yorick CHERN[c], Ramon RIOJAS[a], Omar S. LATIF[d], Usha LOKALA[e], Dexter HADLEY[f], Amit SHETH[e] and Ramin E. BEYGUI[a]
[a] *Department of Surgery, Division of Cardiothoracic Surgery, School of Medicine, University of California San Francisco, San Francisco, CA, USA*
[b] *Bakar Computational Health Sciences Institute, School of Medicine, University of California San Francisco, San Francisco, CA, USA*
[c] *Department of Bioengineering, University of California Berkeley, Berkeley, CA, USA*
[d] *Department of Internal Medicine, Mayo Clinic, Rochester, MN, USA*
[e] *Department of Computer Science and Engineering, Artificial Intelligence Institute, University of South Carolina, Columbia, SC, USA*
[f] *Department of Clinical Sciences, College of Medicine, University of Central Florida, Orlando, FL, USA*

Abstract. The development of an ontology facilitates the organization of the variety of concepts used to describe different terms in different resources. The proposed ontology will facilitate the study of cardiothoracic surgical education and data analytics in electronic medical records (EMR) with the standard vocabulary.

Keywords. Ontology, Cardiothoracic Surgery, Common Vocabulary, Education

1. Introduction

The variety of terms typically employed in cardiothoracic surgical education and study can challenge reasoning over different resources due to other representations. We can overcome these challenges with the standard vocabularies that facilitate various concepts' organization as a general idea of something with an ontology [1]. There are several valuable ontologies and terminologies in a medical domain, such as SNOMED CT[2] (Systematized Nomenclature of Medicine-Clinical Terms), Unified Medical Language System[3] (UMLS), and Cardiovascular Disease Ontology (CVDO)[4]. But, none of the above cover all terms for clinical study in one place with related codes (e.g., ICD10 and Current Procedural Terminology (CPT) codes are not included in ontologies with all corresponding codes).

[1] Corresponding Author, Maryam PANAHIAZAR; E-mail: Maryam.panahiazar@ucsf.edu
[2] http://www.nlm.nih.gov/research/umls/Snomed/snomed_main.html
[3] http://www.nlm.nih.gov/research/umls/
[4] https://bioportal.bioontology.org/ontologies/CVDO

2. Methods

Our systematic approach to the ontology creation is shown in figure 1. In the process of developing the ontology with web protégé [4], we created a new term or imported the term from other resources, we created the definition for each term, and added annotations (e.g., adding comments). Terms are selected from related resources, as shown in figure 1, and with discussion with the experts, we validated the selections.

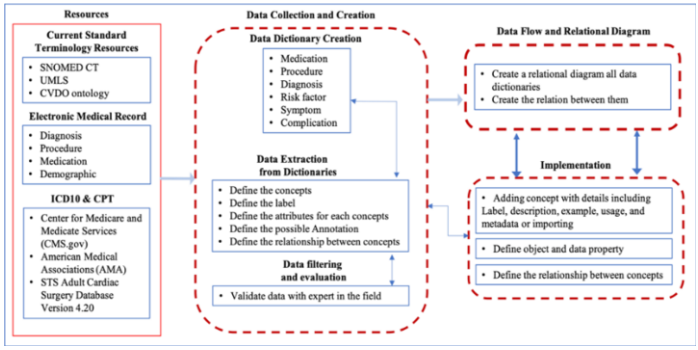


Figure 1. shows a systematic approach to the design and development of CTS ontology.

3. Results and Discussion

Our ontology includes descriptions of classes, definitions, and relations. It includes 1041 medication codes in levels of therapeutic, pharmaceutical class, and subclass, ~210 procedure, and ~200 ICD10 codes, 58 risk factors, and the relationships (e.g., "is-a", "has-a", "as a result of"). For example, CABG "is-a" surgical procedure "to restore" normal blood flow to an obstructed coronary artery. Statin "is-a" cholesterol-lowering drug "is-taken-by" patients with hyperlipidemia (ICD10 = E78.5) "caused-by" (e.g., obesity).

4. Conclusions

We will deploy our ontology within the BioPortal [2] as a web-services-based portal to enable universal accessibility over the Internet. It allows researchers to access, explore and reuse the resources and products of the related domain [3]. The novelty of this work is filling the gap between different resources with complete ontology for surgical education and data analytics that includes related coeds from EMR.

References

[1] Gruber TR. A translation approach to portable ontology specifications. Knowledge acquisition. 1993 Jun 1;5(2):199-220.
[2] Noy NF, et al. BioPortal: ontologies and integrated data resources at the click of a mouse. Nucleic acids research. 2009 Jul 1;37(suppl_2):W170-3.
[3] Panahiazar M, Sheth AP, Ranabahu A, et al. Advancing data reuse in phyloinformatics using an ontology-driven Semantic Web approach. BMC medical genomics. 2013 Nov;6(3):1-0.
[4] Webpage : https://webprotege.stanford.edu, *Update*. (2022).

Challenges of Trustable AI and Added-Value on Health
B. Séroussi et al. (Eds.)

doi:10.3233/SHTI220486

Large-File Raw Data Synchronization for openBIS Research Repositories

Aleksei DUDCHENKO[a,1], Friedemann RINGWALD[a], Felix CZERNILOFSKY[b], Sascha DIETRICH[b], Petra KNAUP[a] and Matthias GANZINGER[a]
[a] *Institute of Medical Informatics, Heidelberg University Hospital, Germany*
[b] *Department of Hematology, Heidelberg University Hospital, Germany*

Abstract. In a systems medicine research consortium, openBIS is used as a research data repository. To facilitate efficient upload of large files, openBIS is complemented by a Nextcloud data cloud system. Using a Nextcloud client, raw mass spectrometry data is automatically imported into the repository in the background, enabling comprehensive data provenance.

Keywords. systems medicine, research platform, mass spectrometry

1. Introduction

SMART-CARE (Systems Medicine Approach to Stratification of Cancer Recurrence) is a research project of several clinical, biomedical, and medical informatics partners [1]. Its goal is to build a robust mass spectrometry (MS)-based pipeline for predicting relapse of cancer patients. For all steps of the pipeline, standard operating procedures are developed to ensure reproducible proteome and metabolome analyses. The requirements of individual lab partners for a complex solution for MS data management led to the SMART-CARE Linked Data Repository (Smart LDR) based on the open-source openBIS software [2] to assure correct sample and experiment identification. In this solution, security and access management are provided for large data sets. To ensure complete reproducibility of the analyses and allow for the application of new analytical methods in the future, the consortium decided to add raw MS data files to the Smart LDR. Since the regular web-based interface of openBIS proved to be inconvenient for sharing up to 90 files per experiment with sizes of three to eight gigabytes each, an extension for automatically synchronizing these files had to be newly developed.

2. Methods

For the Smart LDR, openBIS was adopted for collecting, managing, and exchanging complex data and metadata provided by clinics, laboratories, and analytical partners. The raw files generated by mass-spectrometry equipment have to be uploaded to a data server via a web browser interface and are attached to corresponding objects in the Smart LDR.

[1] Corresponding author, Aleksei DUDCHENKO; E-mail: aleksei.dudchenko@med.uni-heidelberg.de

Users suggested the upload process should be done in the background and should handle possible internet disruptions. The uploaded files should be linked to the corresponding objects in Smart LDR containing annotations without manual interaction.

We researched existing approaches allowing for smooth upload of large files. A class of software that is well-established, are systems for setting up private data clouds [3], such as Seafile or Nextcloud. Further requirements were the existence of a local client program in addition to the web browser interface. Since local firewalls only allow for network access via HTTP proxies, only systems with a client suitable for such an environment were considered.

3. Results

As a result of our research, we decided to use the open-source software Nextcloud. The software has a mature state of development and is complemented by client software that works well with HTTP proxies. Nextcloud was set up on a dedicated virtual server and solves the task of transferring data from laboratory machines to the Smart LDR data store server. This was achieved by defining an external storage device in Nextcloud for each lab. The labs were instructed to install the client software and configure it to automatically synchronize a local drop directory with the respective folder on the server.

The server-side storage devices are linked into the openBIS data storage server via Network File System (NFS). As soon as completely transferred files are detected by openBIS, files are moved to the final storage location and deleted from the local directory. Customized openBIS scripts and rules assure correct assignment between files and samples respectively experiments based on a naming convention of the files.

4. Discussion and Conclusions

The solutions we implemented for MS raw data exchange allow seamless data upload for labs and fully automated data binding to responsible samples and objects in the Smart LDR. The major outcome of openBIS extension is the decrease of workload and reduction of burden for lab staff while at the same time ensuring a complete and error-free raw data repository. Since the concept is generic, our approach can be adopted by other projects using openBIS.

SMART-CARE is funded by the German Ministry of Education and Research (Grant id: 031L0212A).

References

[1] Ringwald F, Czernilofsky F, Dudchenko A, Ganzinger M, Dietrich S, Knaup P. Data Management for Systems Medicine: The SMART-CARE Joint Environment. Stud Health Technol Inform. 2021 May 27;281:1104-1105. doi: 10.3233/SHTI210365. PMID: 34042858.

[2] Bauch A, Adamczyk I, Buczek P, Elmer FJ, Enimanev K, Glyzewski P, Kohler M, Pylak T, Quandt A, Ramakrishnan C, Beisel C, Malmström L, Aebersold R, Rinn B. openBIS: a flexible framework for managing and analyzing complex data in biology research. BMC Bioinformatics. 2011 Dec 8;12:468.

[3] Kanbar LJ, Shalish W, Robles-Rubio CA, Precup D, Brown K, Sant'Anna GM, Kearney RE. Organizational principles of cloud storage to support collaborative biomedical research. Annu Int Conf IEEE Eng Med Biol Soc. 2015 Aug;2015:1231-4. doi: 10.1109/EMBC.2015.7318589. PMID: 26736489.

Challenges of Trustable AI and Added-Value on Health
B. Séroussi et al. (Eds.)

doi:10.3233/SHTI220487

Causal Associations Among Diseases and Imaging Findings in Radiology Reports

Ronnie SEBRO[a] and Charles E. KAHN, Jr.[b,1]
[a] *Mayo Clinic, Jacksonville, Florida, USA*
[b] *University of Pennsylvania, Philadelphia, Pennsylvania, USA*

Abstract. This study explored the ability to identify causal relationships between diseases and imaging findings from their co-occurrences in radiology reports. A natural language processing (NLP) system with negative-expression filtering detected positive mentions of 16,912 disorders, interventions, and imaging findings in 1,702,462 consecutive radiology reports; the 55,564 causal relations defined by the Radiology Gamuts Ontology (RGO) served as reference standard. Conditions were considered to co-occur if they were present in reports from the same patient. The φ and κ statistics both achieved AUC≥0.70, P<0.001 in identifying causal relationships from pairwise co-occurrence data. Analysis of radiology reports can identify a large proportion of known causal associations among diseases and imaging findings. Automated approaches hold promise to identify causal relationships among diseases and imaging findings from their co-occurrence in text-based radiology reports.

Keywords. Knowledge discovery, Ontologies, Health data science, Big data, Natural language processing, Radiology, Reporting

1. Introduction

Narrative-text radiology reports provide a rich source of information that can be extracted using natural language processing (NLP) to identify the presence of diseases and imaging findings. This study applied an ontology of disorders, interventions, and imaging observations to detect co-occurrence of diseases and imaging findings in a large corpus of narrative-text radiology reports. The study evaluated the ability of two statistical tests to identify causal relationships based on the co-occurrence of terms in a large cohort of patients**.**

2. Materials and Methods

1,702,462 consecutive radiology reports of 1,396,293 patients were analyzed in this IRB-approved study. The Radiology Gamuts Ontology (RGO) defined 16,912 imaging findings, diseases, and interventions, and 55,564 causal relationships between them [1]. An occurrence was defined as positive mention of an RGO entity; reports were

[1] Corresponding Author, Charles E. Kahn, Jr, MD, MS, Department of Radiology and Institute for Biomedical Informatics, University of Pennsylvania, 3400 Spruce Street, Philadelphia PA 19104, USA; E-mail: ckahn@upenn.edu.

aggregated by patient. Analysis included pairs of entities that co-occurred in at least 25 patients. The absolute value of the phi coefficient (φ) and Cohen's kappa statistic (κ) were analyzed: receiver operator characteristic (ROC) curves were compared using DeLong's test. Area under the ROC curve (AUC) was computed.

3. Results

Analysis was limited to 95,620 pairs where both RGO entities occurred in 25 or more patients, of which 161 pairs (0.17%) had causal associations in RGO. Both φ and κ achieved AUC $\geq$ 0.70; their AUCs showed no statistically significant difference (DeLong's test p=0.288).

4. Conclusion

The φ and κ statistics showed moderate performance in identifying known causal relationships based on pairwise co-occurrence of terms in radiology reports; there was no statistically significant difference between the two metrics. Both metrics' performance can be considered remarkably strong given that "positive" pairs of causally related entities constituted only 0.17% of all of pairs of entities evaluated. This finding suggests that automated approaches may be able to help identify causal relationships among diseases and imaging findings from text-based radiology reports. Our findings agree with reports of pairwise disease-finding associations in clinical data and have mined observational data to identify causal pathways [2, 3]. In most settings, most EHR information exists as narrative text, including radiology reports.

Our larger goal is to discover new causal relationships among disorders and imaging findings directly from their patterns of co-occurrence. Discovery of causal relationships in observational data has been grounded in Bayesian network learning, a computationally intensive (and NP-complete) problem; various approaches have sought to reduce the computational burden to make the approach more tractable on large datasets [4, 5]. Given our large cohort of observational data, our current work includes induction of a Bayesian network model over the entire set of co-occurring terms. Preliminary analysis has shown significantly stronger performance than pairwise co-occurrence statistics and has yielded plausible new causal relations.

References

[1] Budovec JJ, Lam CA, Kahn CE, Jr. Radiology Gamuts Ontology: differential diagnosis for the Semantic Web. RadioGraphics. 2014;34(1):254-64.
[2] Cao H, Hripcsak G, Markatou M. A statistical methodology for analyzing co-occurrence data from a large sample. J Biomed Inform. 2007;40(3):343-52.
[3] Cao H, Markatou M, Melton GB, Chiang MF, Hripcsak G. Mining a clinical data warehouse to discover disease-finding associations using co-occurrence statistics. AMIA Annu Symp Proc. 2005:106-10.
[4] Jin Z, Li J, Liu L, Le TD, Sun B, Wang R, editors. Discovery of causal rules using partial association. IEEE 12th International Conference on Data Mining; 2012 10-13 Dec. 2012.
[5] Ramanan N, Natarajan S. Causal learning from predictive modeling for observational data. Front Big Data. 2020;3:535976.

Challenges of Trustable AI and Added-Value on Health
B. Séroussi et al. (Eds.)

doi:10.3233/SHTI220488

Beyond the Brain: MIDS Extends BIDS to Multiple Modalities and Anatomical Regions

Jose Manuel SABORIT-TORRES[a,*], Silvia NADAL-ALMELA[a,*,1], Joaquim Angel MONTELL-SERRANO[a], Elena OLIVER-GARCIA[a], Hector CARCELLER[a,b,c], Jon Ander GÓMEZ-ÁDRIAN[d], Marisa CAPARRÓS-REDONDO[a], Francisco GARCÍA-GARCÍA[a,e], Julio DOMENECH-FERNÁNDEZ[f] and Maria DE LA IGLESIA-VAYÁ[a,b,g]

[*] *Equal contributors*

[a] *Unidad Mixta de Imagen Biomédica FISABIO-CIPF. Fundación para el Fomento de la Investigación Sanitaria y Biomédica de la Comunidad Valenciana – València, Spain*

[b] *CIBERSAM, ISC III. Av. Blasco Ibáñez 15, 46010 - València, Spain*

[c] *Institute for Biotechnology and Biomedicine (BIOTECMED), Universitat de Valencia, Burjassot, 46100, Spain*

[d] *Department of Computer Systems and Computation, Universitat Politècnica de València - València, Spain*

[e] *Bioinformatics & Biostatistics Unit, Principe Felipe Research Center – Valencia, Spain*

[f] *Hospital Arnau de Vilanova – Valencia, Spain*

[g] *General Directorate of Research, Innovation, Technology and Quality. Subdirectorate General of Information Systems for Health – Valencia, Spain*

Abstract. Brain Imaging Data Structure (BIDS) provides a valuable tool to organise brain imaging data into a clear and easy standard directory structure. Moreover, BIDS is widely supported by the scientific community and has been established as a powerful standard for medical imaging management. Nonetheless, the original BIDS is restricted to magnetic resonance imaging (MRI) of the brain, limiting its implantation to other techniques and anatomical regions. We developed Medical Imaging Data Structure (MIDS), conceived to extend BIDS methodology to other anatomical regions and multiple imaging systems in these areas. The MIDS standard was developed to store and manage medical images as an extension of BIDS. It allows the user to handily save studies of multiple anatomical regions and imaging techniques. Besides, MIDS improves the classification of multiple images within the structure, allowing the possibility to unify them in a single study to apply on them preprocessing or artificial intelligence algorithms. Finally, the results generated are saved in the derivatives folder.

Keywords. Database, BIDS, Standardization, OMOP

[1] Corresponding Author, Silvia Nadal Almela; E-mail: s.nadalm@gmail.com.

1. Introduction

Brain Imaging Data Structure (BIDS) has been established as a powerful standard to organize brain imaging data and is widely supported by the scientific community. In order to extend BIDS to other anatomical regions beyond the brain, we have developed Medical Imaging Data Structure (MIDS).

2. Methods and Resutls

BIDS [1] is a standard for storing magnetic resonance imaging data and metadata in a clear and simple hierarchical folder structure. It is supported by several programs and libraries dedicated to the study of medical images (e.g., c-pacs, freesurfer, XNAT, BIDS Validator, among others) and is widely used by research groups.

MIDS [2] expands BIDS structure by including a general template for other anatomical regions (chest, pelvis, prostate, spinal...) and file tags for describing many aspects of the data such as body part, view position, DICOM modality.

3. Conclusions and Future Work

MIDS improves the structure categorization and usability for the user providing a common structure for many projects, which can be helpful in artificial intelligence projects. MIDS is part of a collaboration with the DeepHealth EU project in the guidance and construction of a fully anonymized population medical imaging data lake structure. As part of project TARTAGLIA, a new extension of MIDS is being implemented to make it compatible with the Observational Medical Outcomes Partnership (OMOP) Common Data Model [3].

References

[1] Gorgolewski KJ, Alfaro-Almagro F, Auer T, Bellec P, Capotă M, Chakravarty MM, Churchill NW, Cohen AL, Craddock RC, Devenyi GA, Eklund A, Esteban O, Flandin G, Ghosh SS, Guntupalli JS, Jenkinson M, Keshavan A, Kiar G, Liem F, Raamana PR, Raffelt D, Steele CJ, Quirion PO, Smith RE, Strother SC, Varoquaux G, Wang Y, Yarkoni T, Poldrack RA. BIDS apps: Improving ease of use, accessibility, and reproducibility of neuroimaging data analysis methods. PLoS Comput Biol. 2017 Mar 9;13(3):e1005209. doi: 10.1371/journal.pcbi.1005209. PMID: 28278228; PMCID: PMC5363996.

[2] Sabori-Torres JM, Saenz-Gamboa JJ, Montell JÀ, Salinas JM, Gómez JA, Stefan I, Caparrós M, García-García F, Domenech J, Manjón JV, Rojas G. Medical imaging data structure extended to multiple modalities and anatomical regions. arXiv preprint arXiv:2010.00434. 2020 Sep 30.

[3] Observational Health Data Sciences and Informatics (OHDSI). https://www.ohdsi.org/

Challenges of Trustable AI and Added-Value on Health
B. Séroussi et al. (Eds.)

doi:10.3233/SHTI220489

The Common Provenance Model: Capturing Distributed Provenance in Life Sciences Processes

Francesca FREXIA[a,1], Cecilia MASCIA[a], Rudolf WITTNER[b,c], Markus PLASS[d], Heimo MÜLLER[d], Jörg GEIGER[e] and Petr HOLUB[b]
[a] *CRS4 – Center for Advanced Studies, Research and Development in Sardinia, Italy*
[b] *BBMRI-ERIC, Austria*
[c] *Masaryk University, Czech Republic*
[d] *Diagnostic and Research Institute of Pathology, Medical University of Graz, Austria*
[e] *Interdisciplinary Bank of Biomaterials and Data Würzburg (ibdw), Germany*

Abstract. The distributed nature of modern research emphasizes the importance of collecting and sharing the history of digital and physical material, to improve the reproducibility of experiments and the quality and reusability of results. Yet, the application of the current methodologies to record provenance information is largely scattered, leading to silos of provenance information at different granularities. To tackle this fragmentation, we developed the Common Provenance Model, a set of guidelines for the generation of interoperable provenance information, and to allow the reconstruction and the navigation of a continuous provenance chain. This work presents the first version of the model, available online, based on the W3C PROV Data Model and the Provenance Composition pattern.

Keywords. Provenance information, distributed processes, W3C PROV, Provenance Composition, Common Provenance Model

1. Introduction

Knowing the origin and evolution of data and samples is crucial to assess and improve the reliability and reusability of the results generated from their processing, as recognised also by the FAIR initiative [1]. Some approaches to express provenance information (PI) exist, but the lack of an overall coordination severely limit their impact, especially in *distributed* processes (e.g., large-scale analysis, AI applications), generally involving heterogeneous and complex steps, often performed by several groups at different times. We present here the Common Provenance Model (CPM) [2], a set of methodological recommendations we developed to guide the creation and exchange of interoperable PI.

2. Methods

The CPM design followed a domain-agnostic approach, considering different research processes in the life sciences within the EOSC-Life Project (involving thirteen European research infrastructures in the life sciences) and regional initiatives in Sardinia. The

[1] Corresponding Author, Francesca Frexia, CRS4, Italy; E-mail: francesca.frexia@crs4.it.

model focuses on three fundamental dimensions of the provenance lifecycle: the formalisation, linking, and navigation of a chain of provenance information. For the generation of provenance descriptions, we have adopted the W3C PROV formalism [3], expressing the provenance via instances formed by simple structures (*entity, agent, activity*) linked to relevant domain ontologies, to capture accurately the semantic of each use case. To support linking and navigation of PI in a distributed chain, we revised the Provenance Composition, a conceptual methodology connecting the PI records of two communicating processes through a shared entity [4]. The CPM extends the Provenance Composition approach to couple adjacent tasks, by defining practical recommendations for the PI navigation.

3. Results and Conclusions

The Common Provenance Model prescribes that the PI to be stored or further shared is generated in the form of PROV instances during *finalisation events*, periodically or on request, from the combination of all the relevant details recorded during the execution of a process (e.g., logs, metadata, adopted protocols). The transfer of "objects" – samples or data – is represented in the model introducing a specific PROV entity, the *connector*, to link the sender and the receiver through URIs univocally assigned. The model includes guidelines about fundamental elements to generate PI of good quality, like a versioning mechanism or the management of persistent identifiers. The first version of the CPM is available online [2], it has not been experimentally tested yet, but this limitation is mitigated by the fact that its development was guided by several use cases. Next steps include the refinement and application of the model. To encourage the adoption in the industrial sector, an instance of the CPM is the core of the *"ISO 23494: Biotechnology - Provenance Information Model for Biological Specimen and Data"* project [5], currently under development. Experts in the life sciences are invited to collaborate, through a national ISO body or by engaging with EOSC-Life online meetings.

Author Contributions and Acknowledgements

R.W. was the primary author of the distributed provenance information model. All the other authors provided continuous feedback and worked on refinements of the initial draft. This poster was prepared by F.F. and C.M. and the rest of the authors revised it. The work described is being co/funded by EOSC-Life Project (EU Horizon 2020, grant agreement no. 824087) and by DIFRA, SVDC and PAM Projects (Sardinian Regional Authority). This publication reflects only the view of the authors, and the European Commission is not responsible for any use that may be made of the information it contains.

References

[1] Wilkinson M, et al. The FAIR Guiding Principles for scientific data management and stewardship. Scientific data. 2016;3(1):1-9.
[2] Wittner R, et al. EOSC-Life Common Provenance Model. Zenodo; 2021
[3] Groth P, Moreau L. PROV-Overview: An Overview of the PROV Family of Documents. (Accessed on 01/19/2022). https://www.w3.org/TR/prov-overview/.
[4] Buneman P, et al. Provenance Composition in PROV; 2017. https://eprints.soton.ac.uk/408513/
[5] Wittner R, et al. Provenance Week 2020 Poster: ISO 23494: Biotechnology - Provenance Information Model for Biological Specimen and Data. Zenodo; 2020.

Challenges of Trustable AI and Added-Value on Health
B. Séroussi et al. (Eds.)

doi:10.3233/SHTI220490

Implementation of Gene Expression Profiles in the HL7 FHIR Standard

Florian AUER[a,1], Zhibek ABDYKALYKOVA[a], Dominik MÜLLER[a] and Frank KRAMER[a]
[a] *IT-Infrastructure for Translational Medical Research, University of Augsburg, Germany*

Abstract. Gene expression profiles can capture significant molecular differences paving the way toward precision medicine. However, clinical standards like FHIR only provide encoding of molecular sequence variations, even so, expression patterns are equally important for decision making. Here we provide an exemplary implementation of gene expression profiles of a microarray analysis using an adaption of the FHIR Genomics extension. Our results demonstrate how FHIR resources can be facilitated in bioinformatics-based decision support systems or used for the aggregation of molecular genetics data in multi-center clinical trials.

Keywords. FHIR, interoperability, omics, gene expression

1. Introduction

Over the years, interoperability gained more importance in clinical settings with standards like Fast Healthcare Interoperability Resources (FHIR) [1]. However, the genomic profiles in the FHIR Genomics extension only cover variations in the molecular sequence while expression patterns are neglected. Nevertheless, insights from expression patterns are important for decision support and translational medical research. Here we present a feasible FHIR implementation for gene expression profiles.

2. Methods

The central element within FHIR to capture real-world concepts in healthcare systems is the *Patient* resource, which is why all subsequent patient-specific results and resources were referred to this base. For the preservation of the anonymity of participants, we used artificially generated data to create *Patient* resources using Synthea™ [2]. The medical condition was captured by the *Condition* and samples by the *Specimen* resource to allow multiple samples for a single patient. Referenced genes were included as *MolecularSequence* resources, which avoids redundancy in common genomic profiles and simplifies retrieval of expression values for a single gene across multiple samples. The expression values are treated as single measurements of *Observation-geneticsGene* resources referring to the corresponding gene. An overview is shown in figure 1.

[1] Corresponding Author,s Florian Auer, IT Infrastructure for Translational Medical Research, Alter Postweg 101, 86159 Augsburg, Germany; E-mail: florian.auer@informatik.uni-augsburg.de

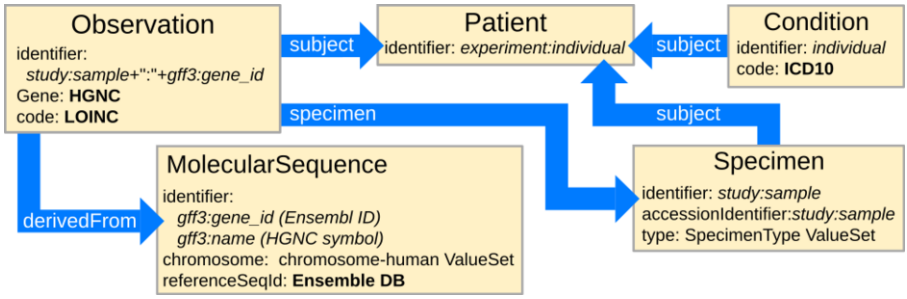


Figure 1. Overview of the extended FHIR resources and references for gene expression profiles.

3. Results

For demonstration, we utilized a HAPI FHIR server [3] and an mRNA microarray [4], which examines a dose-limiting side effect of chemotherapy in patients with acute myeloid leukemia (EBI Expression Atlas: E-GEOD-10746). The data set translated to 252,684 resources stored on our FHIR server. Our implemented FHIR solution is available on GitHub: https://github.com/frankkramer-lab/gene-expression-on-fhir.

4. Discussion

Through our contribution to the FHIR Genomics extension, we were able to include genomic profiling data allowing enhanced clinical interoperability of genomic data. We are planning further evaluation by application in a decision support system as well as consolidation of the outcome of expression analyses as *DiagnosticReport* resources.

5. Conclusions

Our results demonstrate how FHIR resources can be facilitated for the clinical exchange of expression profiles. The further extension of FHIR allows for establishing a currently missing standard for molecular genetics data in clinical settings.

References

[1] Vázquez SG, Barreira N, Penedo MG, Pena-Seijo M, Gómez-Ulla F. Evaluation of SIRIUS retinal vessel width measurement in REVIEW dataset. InProceedings of the 26th IEEE International Symposium on Computer-Based Medical Systems 2013 Jun 20 (pp. 71-76). IEEE.

[2] Walonoski J, et al. Synthea: An approach, method, and software mechanism for generating synthetic patients and the synthetic electronic health care record. J Am Med Inform Assoc. 2018 Mar 1;25(3):230-238. doi: 10.1093/jamia/ocx079. Erratum in: J Am Med Inform Assoc. 2018 Jul 1;25(7):921.

[3] Sánchez YK, Demurjian SA, Baihan MS. Achieving RBAC on RESTful APIs for mobile apps using FHIR. In2017 5th IEEE International Conference on Mobile Cloud Computing, Services, and Engineering (MobileCloud) 2017 Apr 6 (pp. 139-144). IEEE.

[4] Mougeot JL, Bahrani-Mougeot FK, Lockhart PB, Brennan MT. Microarray analyses of oral punch biopsies from acute myeloid leukemia (AML) patients treated with chemotherapy. Oral Surg Oral Med Oral Pathol Oral Radiol Endod. 2011 Oct;112(4):446-52.

Challenges of Trustable AI and Added-Value on Health
B. Séroussi et al. (Eds.)

doi:10.3233/SHTI220491

REDIRECT: Mapping Drug Prescriptions and Evidence from Biomedical Literature

Bruno THIAO LAYEL[a,b,1], Vianney JOUHET[a,c], Guillaume BLIN[b] and Gayo DIALLO[a,b]
[a] *BPH Center - INSERM U1219, Team AHeaD, Univ. Bordeaux, Bordeaux, France*
[b] *LaBRI - CNRS UMR5800, Team BKB, Univ. Bordeaux, Talence, France*
[c] *CHU de Bordeaux, Pôle de santé publique, SIM, unit IAM, Bordeaux, France*

Abstract. To enhance their practice, healthcare professionals need to cross-link various usage recommendations provided by heterogeneous vocabularies that must be retrieved and integrated conjointly. This is the aim of the Knowledge Warehouse / K-Ware platform. It enables establishing relevant bridges between different knowledge sources (structured vocabularies, thesaurus, ontologies) expressed in the semantic web standard languages (i.e. SKOS, OWL, RDF). This poster presents the strategy applied in K-Ware to hide the different aspects of linking literals with medical entities encoded in these knowledge sources to fetch some publications abstracts from Pubmed.

Keywords. Knowledge Resource, RDF, meta-model, ontology, terminology, semantic, Knowledge warehouse

1. Introduction

The poster focuses on drug prescribing, which is an old known problem in Public Health Informatics, using Pubmed as a supplementary resource providing more up to date therapeutic advances. We tackle the integration of heterogeneous Semantic Knowledge Resources **SKR** at several levels by their format but also by their representation within vocabularies available in the Linked Open Data [1] **LOD**. Its life-science part is full of SKRs about *symptoms, indications, drugs* and so on, persisted and defined by standardized vocabularies such as **RDF** [2], **OWL** [3] or **SKOS** [4]. These schemes are natural tools to bring meaning to concrete data that would not be linked at first sight.

2. Methods

We have to seek external knowledge in order to refine a **MeSH** (Medical Subject Heading [5]) query to be performed on Pubmed. The Romedi resource [6] provides us enough metadata by structuring the open Public Drug Database and marketed in France [7] in an RDF format. Romedi provides links between a marketed drug and its therapeutic indications from the **ATC** (Anatomical Therapeutic Chemical Classification [8]) and

[1] Corresponding Author, Bruno THIAO LAYEL; E-mail: bruno.thiao-layel@u-bordeaux.fr

also their bibliographic tag to MeSH terms themselves represented as SKOS-based terminologies.
K-Ware is the framework used for handling each of the RDF-based SKR, provides ways to:

1. *Recognise* a drug name from a Literal as an identified entity within Romedi
2. *Browse* a marketed drugs' attributes from Romedi to retrieve its active substance as an ATC entity
3. *Handle mappings* between MeSH terms and ATC codes

Then establishing a cooccurrence query with all the MeSH terms retrieved from a starting prescription's expanded entities in order to retrieve PubMed's abstracts.

3. Results

An implemented GUI is available to specify a prescription order by using a drug/active substance name auto-complete.
When a prescription for at least 2 drugs is formed, the interface displays the different publication's abstract fetched from Pubmed.
The scenario is testable at *http://kware.erias.fr*.

4. Conclusions

K-Ware being more considered as an RDF framework rather than a simple termino-ontological resource repository, allows by using its internal metamodel, the heterogeneous management of the Romedi vocabulary which is an RDF database along with the ATC and MeSH vocabularies which are rather SKOS terminologies without having to modify them.
The K-Ware framework, by allowing us to move towards less data, but more metadata provides ways to not distort any represented Semantic Knowledge Resources for a well-defined project context.

References

[1] Bauer F, Kaltenböck M. Linked open data: The essentials. Edition mono/monochrom, Vienna. 2011;710.
[2] RDF 1.1 Concepts and Abstract Syntax, [Online; Accessed: 2022-03-31].
[3] W3C, OWL 2 Web Ontology Language Document Overview (Second Edition), [Online; accessed 2022-03-31].
[4] W3C, SKOS Simple Knowledge Organization System Reference, [Online; accessed 2022-03-31].
[5] Lipscomb CE. Medical subject headings (MeSH). Bulletin of the Medical Library Association. 2000 Jul;88(3):265.
[6] Cossin S, Lebrun L, Lobre G, Loustau R, Jouhet V, Griffier R, Mougin F, Diallo G, Thiessard F. Romedi: An Open Data Source About French Drugs on the Semantic Web. Stud Health Technol Inform. 2019 Aug 21;264:79-82. doi: 10.3233/SHTI190187. PMID: 31437889.
[7] Accueil - Base de données publique des médicaments, [Online; Accessed: 2022-03-31].
[8] Skrbo A, Begović B, Skrbo S. Klasificiranje lijekova po ATC sistemu i najnovije izmjene [Classification of drugs using the ATC system (Anatomic, Therapeutic, Chemical Classification) and the latest changes]. Med Arh. 2004;58(1 Suppl 2):138-41. Bosnian. PMID: 15137231.

Challenges of Trustable AI and Added-Value on Health
B. Séroussi et al. (Eds.)

doi:10.3233/SHTI220492

Ontology-Driven Data Cleaning Towards Lossless Data Compression

Athanasios KIOURTIS[a,1], Argyro MAVROGIORGOU[a], George MANIAS[a] and Dimosthenis KYRIAZIS[a]
[a] *Department of Digital Systems, University of Piraeus, Greece*

Abstract. With the available data in healthcare, healthcare organizations and practitioners require interoperable, efficient, and non-time-consuming data exchange. Currently, several cases aim to the exchanged data security, without considering the complexity of the data to be exchanged. This paper provides an Ontology-driven Data Cleaning mechanism, facilitating Lossless Healthcare Data Compression to efficiently compress healthcare data of different nature (textual, audio, image). The latter is being evaluated considering three datasets of different formats, concluding to the added value of the described mechanism.

Keywords. Ontologies, Lossless Compression, Health Information Exchange

1. Introduction

With the devices' increase, there has been a growth of data exchange protocols. Especially in healthcare, due to the broader eHealth strategy [1], healthcare organizations require interoperable data and efficient exchange. Current approaches pay attention to the exchanged data security, without fully facing Health Information Exchange (HIE) requirements [2], due to non-efficient transfer rates and low performance data exchange protocols. Data compression algorithms provide efficient rates, which may lead to information loss, driving the implementation of lossless compression algorithms [3].

In this paper, it is being introduced an Ontology-driven Data Cleaning mechanism, facilitating Lossless Healthcare Data Compression, which parses and splits healthcare data into data chunks, and based on the ingested data type and format, different lossless compression algorithms are provided to function in parallel threads.

2. Methods

The Ontology-driven Data Cleaning mechanism for Lossless Healthcare Data Compression follows a three-stepped process (Fig. 1). The 1st step (*XML Parsing*) [4] parses the data, creating its ontological structure. The 2nd step (*Ontology Containerization*) extracts the skeleton of the data (Structure Container) and the data content per se (Data Container) and splits it into data chunks. The 3rd step (*Data*

[1] Corresponding Author, Athanasios Kiourtis, 80, M. Karaoli & A. Dimitriou St., 18534 Piraeus, Greece; E-mail: kiourtis@unipi.gr.

Encoding) compresses the data, following Lempel–Ziv–Markov chain (LZMA) for texts, DEFLATE for images, and MPEG-4 Audio Lossless Coding (ALS) for audio [3].

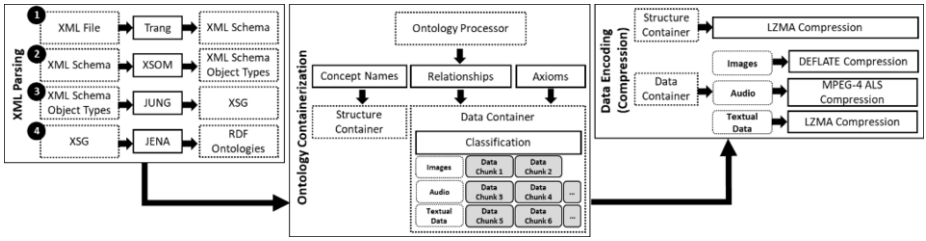


Figure 1. Architecture of the Ontology-driven Lossless Healthcare Data Compression.

3. Results, Evaluation and Discussion

Three different sized datasets have been used, with different data types (image, audio, text), to evaluate the mechanism. The overall process provided the insights of Table 1.

Table 1. Results of the Ontology-driven Data Cleaning towards Lossless Data Compression mechanism.

Dataset Number	Initial Overall Size (Mb)	Compressed size (Mb)	Compression Rate	Compression time (sec)
#1	8	6.7	16%	1
#2	172	94.3	55%	9
#3	1412	567.2	60%	66

From Table 1, the compression time for the small dataset (#1) was close to zero. However, from the compression time of larger datasets (#2, #3), it can be seen that lossless compression should be used in large-sized datasets, since their transmission rate would be efficiently decreased, facilitating the overall HIE.

What is currently within our next goals is the evaluation of the mechanism with more communication technologies [5] (Wi-Fi, D2D protocol), to identify the most suitable technology for HIE, re-using and adjusting additional lossless compression algorithms.

Acknowledgment

The research leading to the results presented in this paper has received funding from the EU funded Project PolicyCLOUD (GA no 870675) and InteropEHRate (GA no 826106).

References

[1] National eHealth Strategy Toolkit, https://www.who.int/ehealth/publications/overview.pdf
[2] Kiourtis A, Mavrogiorgou A, Kyriazis D, Graziani A, Torelli F. Improving Health Information Exchange through Wireless Communication Protocols. In 16th WiMob. 2020:32-39.
[3] Gupta A, Bansal A, Khanduja V. Modern lossless compression techniques: Review, comparison and analysis. In 2nd ICECCT. 2017:1-8.
[4] Kiourtis A, et al. Aggregating the syntactic and semantic similarity of healthcare data towards their transformation to HL7 FHIR through ontology matching. In IJMI. 2019;132.
[5] Kiourtis A, et al. A Secure Protocol for Managing and Sharing Personal Healthcare Data. In Integrated Citizen Centered Digital Health and Social Care. 2020;92-96.

Section IV

Decision Support Systems and Patient Records

This page intentionally left blank

Challenges of Trustable AI and Added-Value on Health
B. Séroussi et al. (Eds.)

doi:10.3233/SHTI220494

Towards a Clinical Decision Support System for Helping Medical Students in Emergency Call Centers

Edouard MICHOT[a,1*], Jules WOO[a,1*], Louis MOULINE[a,1,*], Catherine SINNAPPAN[a,1*], Adrien BOUKOBZA[a], Florence CAMPEOTTO[a,b,c], Laurent DUPIC[b], Anita BURGUN [a,d,e,f], Benoît VIVIEN[a,b], and Rosy TSOPRA [a,d,e,f]
[a]*Digital Health Program of Université de Paris, Paris, France*
[b]*Régulation Régionale Pédiatrique, SAMU de Paris, Hôpital Necker - Enfants Malades, AP-HP, Paris, France*
[c]*Faculté de Pharmacie, Université de Paris, Inserm UMR S1139, Paris, France*
[d]*Université de Paris, Inserm, Sorbonne Université, Centre de Recherche des Cordeliers, F-75006 Paris, France*
[e]*HeKA, Inria Paris, France*
[f]*Department of Medical Informatics, Hôpital Européen Georges-Pompidou, AP-HP, Paris, France*
**Co-first authors, as they contributed equally to the work*

Abstract. In critical situations such as pandemic, medical students are often called to help in emergency call centers. However, they may encounter difficulties in phone triage because of a lack of medical skills. Here, we aim at developing a Clinical Decision Support System for helping medical students in phone call triage of pediatric patients. The system is based on the PAT (Pediatric Assessment Triangle) and local guidelines. It is composed of two interfaces. The first allows a quick assessment of severity signs, and the second provides recommendations and additional elements such as "elements to keep in mind" or "medical advice to give to patient". The system was evaluated by 20 medical students, with two fictive clinical cases. 75% of them found the content useful and clear, and the navigation easy. 65% would feel more reassured to have this system in emergency call centers. Further works are planned to improve the system before implementation in real-life.

Keywords. Clinical Decision Support System, Pediatric Emergency, Digital Health, Phone call triage.

1. Introduction

Emergency call centers are a key component to regulate patient flow and emergency department crowding [1,2]. They are composed of phone operators who manage patients efficiently and with minimal delays as patients move through stages of care [3]. Phone operators have to assess patients remotely, determine the level of emergency and then decide quickly on the most appropriate medical assistance for the patient [3]. Since recent

[1] Students of the Digital Health Program, Université de Paris; E-mail: edouard.michot@gmail.com, jules.woo@etu.u-paris.fr, louismouline@gmail.com, catherine.sinnappan@etu.u-paris.fr

years, the number of calls is continuously increasing in emergency call centers, while the number of phone operators stabilizes [1]. Some catastrophic events (e.g., COVID-19) may also worsen the situation and lead to an overflow of emergency hotlines. In such critical situations, undergraduate medical students are often involved in phone call triage [4,5]. However, they encounter difficulties because of the lack of medical skills.

Clinical decision supports systems (CDSS) could support undergraduate medical students in telephone triage (TT) [3]. Indeed, CDSS-integrated-TT have the potential to improve patient triage by standardizing the decision-making process independently of operators' qualifications or experience [3]. Several CDSS-integrated-TT have been developed, such as "SALOMON" for nurse triage in Belgium [6] ; "TAS", "Personal Health Adviser", "Centramax", and "AXA Assistance" in UK NHS Direct Call center [7] ; or ExpertRN in US [8]. However, these CDSS are hardly transposable from one setting to another, as it depends strongly on the health care organization of country [3].

In France, emergency call centers lack such CDSS, especially for supporting medical students to manage pediatric emergency situations. Here, we aim at developing a CDSS for supporting undergraduate medical students in patient remote triage of pediatric patients.

2. Methods

2.1 CDSS design

To design the CDSS, we followed 5 steps. ***In step 1***, we made a literature review to identify which guidelines used for building the knowledge base of the CDSS. This resulted in selecting local guidelines written by the SAMU of Paris (Service d'Aide Médicale Urgente, one of the greatest emergency call centers in France), but also the Pediatric Assessment Tool (PAT), an internationally accepted tool for the initial assessment of children [9]. ***In step 2***, we extracted decision and action variables from SAMU guidelines and PAT. These variables were then grouped by categories, depending on their type and their importance in the decision process (e.g., variables of the PAT were grouped into the category "severity symptoms to check in first"). ***In step 3***, we designed several interface mock-ups, using Whimsical platform (https://whimsical.com/, access 2021), for displaying the categories of information identified in step 2, until finding the most appropriate interface. ***In step 4***, we implemented the interface and the underlying algorithm, using Shiny extension of R programming language. ***In step 5***, the prototype was tested by 4 medical doctors from SAMU of Paris, and 7 undergraduate medical students from the digital health program of the Université de Paris. This evaluation resulted in improving the prototype (e.g., adding a more meaningful cursor for the selection of the clinical situation), and led to a new version of the CDSS.

2.2 CDSS evaluation

The CDSS was assessed by 20 undergraduate medical students. They tested the system online without supervision, using two clinical cases simulating emergency phone calls. These clinical cases allowed to navigate through various options of the CDSS. After that, medical students had to rate their experience using an anonymous online form including 6 questions related to aesthetics, relevance and confidence. Each item was assessed through a 5-points Likert Scale. Feedbacks could be added in free-text comment sections.

3. Results

3.1 CDSS design

The CDSS displays two interfaces. The ***first interface*** is dedicated to the quick assessment of pediatric patient. It includes two sections (Figure 1):

- 1st section, on the left (area a), allows to add data on gender, age and weight. If the weight is not provided, then an interactive cursor predicts automatically the most likely weight adjusted to the age (based on pediatric standards [10]).
- 2nd section, on the right, has 2 areas. Top area (area b) specifies the type of emergency (e.g., fever). Bottom area (area c) displays 4 quadrants corresponding to the components of the PAT (i.e., respiratory, circulatory, neurological failures), and the severity signs specific to each type of emergency.

Once the user has filled the first interface, he can visualize the ***second interface***, which displays the recommendations of the CDSS. It includes two sections:

- The 1st section on the left, is the same than in the first interface. It allows the medical student to always keep an eye on patient information.
- The 2nd section on the right, displays the recommended actions and some important medical information. Recommended actions (e.g., sending an ambulance +/- with a rescuer team) are displayed at the top. Additional guidelines and advice are also displayed on request through the tab "medical advice to give to the patient", and the tab "to keep in mind". Finally, a summary of all ticked information is displayed on the bottom, with the possibility to add information in free text. In the future, this recap could be sent to the emergency team on ambulance.

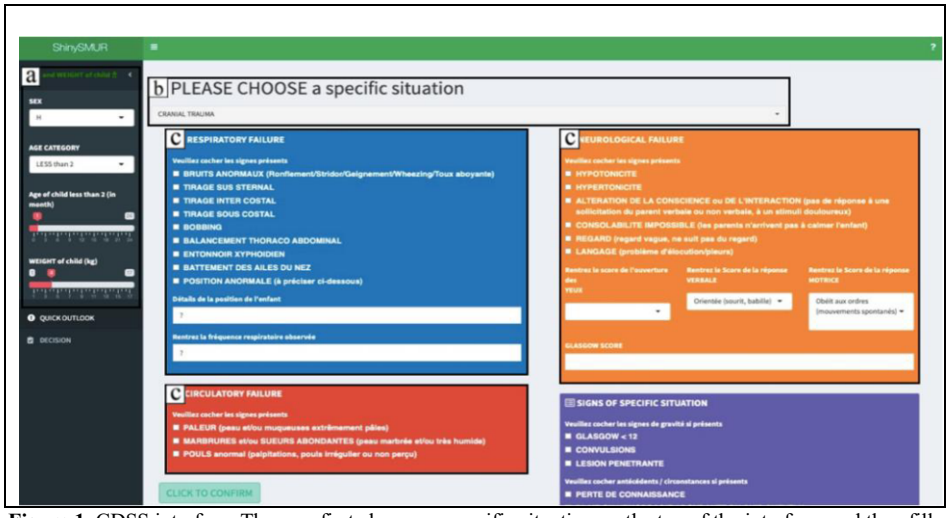


Figure 1. CDSS interface. The user first chooses a specific situation on the top of the interface, and then fills patient information on the left (e.g., age, weight). Then, the user selects the patient symptoms from the list of items. Items are grouped by colour according to the type of failure (e.g., red colour for circulatory failure). Additional items, specific to each clinical situation, are displayed in the purple box.

3.2 CDSS evaluation

Twenty undergraduate medical students assessed the system. 90% of them had no experience in phone call triage, and 40% used medical applications at least once a week.

Regarding user interface, 75% of students found the interface easy to use, 60% liked the design, and 75% found the information clear and useful (Figure 2). Regarding confidence, 65% of students said they would feel more reassured with this system in emergency call centers (30% were undecided) and 45% would have more trust in their decision using this system (45% were undecided). Regarding system use, 55% of students said they would use the system in emergency call centers (40% were undecided).

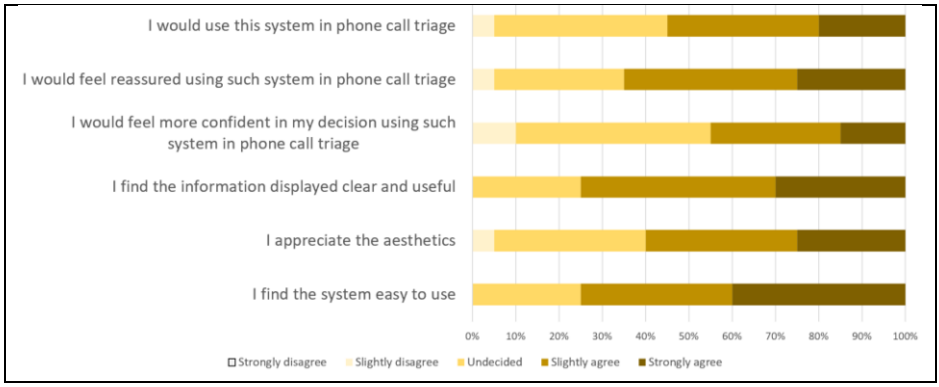


Figure 2. Results of the evaluation.

4. Discussion

We designed a CDSS for helping undergraduate medical students in emergency call centers. The CDSS was designed by and for medical students, with two main objectives: overcoming their lack of expertise in pediatric phone triage and reassuring them in their medical decision. 75% of students found the information useful, and the navigation easy. However, only 45% said they would feel more confident using this system (45% were undecided), and only 55% said they would use the system in emergency call centers (40% were undecided). The high percentage of uncertainty is probably explained by the lack of experience in phone call triage for both children and adults. Indeed, 90% of them had never experienced this task, and therefore may encounter difficulty to determine if the system would reassure them or if they would use it in real life.

This work presents some limits. First, we used a combination of local guidelines and PAT for building the knowledge base, because of a lack of national pediatric guidelines. However, local guidelines were written by a group of medical experts having experienced in a large emergency call center, and PAT is an international well-recognized tool [9]. Other types of representation such as ontologies, could be used in the future to facilitate guidelines implementation. Second, the interface may appear too complex because of high quantity of text (due to our will to provide a maximum of information to medical students to overcome their lack of medical skills). In the future we plan to use other design techniques such as icons for limiting the quantity of text [11].

A few CDSS have been designed for phone call triage [6–8]. However, they targeted specific health professionals (e.g., nurses [6]), and specific organization (e.g., NHS [7]).

Our literature review only retrieved one CDSS designed for medical students, but it was devoted to COVID-19 [4]. We expect that our system could help to standardize medical decision in pediatric phone triage [3], and thus improve the quality of care.

In the future, we aim at improving our system, and then conduct a more robust evaluation, before potential implementation in emergency call center. We expect that the improved system could assist medical students in making quick and appropriate decisions, as well as increasing their medical knowledge in pediatric situations.

Funding

This work is part of an innovative program named "Artificial Intelligence for Clinical Decision Support", dedicated to undergraduate medical students, at Université de Paris.

References

[1] Viglino D, Vesin A, Ruckly S, Morelli X, Slama R, Debaty G, et al. Daily volume of cases in emergency call centers: construction and validation of a predictive model, Scand J Trauma Resusc Emerg Med. 25 (2017) 86. doi:10.1186/s13049-017-0430-9.

[2] Campbell JL, Fletcher E, Britten N, Green C, Holt TA, Lattimer V, et al. Telephone triage for management of same-day consultation requests in general practice (the ESTEEM trial): a cluster-randomised controlled trial and cost-consequence analysis, The Lancet. 384 (2014) 1859–1868. doi:10.1016/S0140-6736(14)61058-8.

[3] Islam F, Sabbe M, Heeren P, Milisen K. Consistency of decision support software-integrated telephone triage and associated factors: a systematic review, BMC Medical Informatics and Decision Making. 21 (2021) 107. doi:10.1186/s12911-021-01472-3.

[4] Vrel JP, Oulmane S, Boukobza A, Burgun A, Tsopra R. A COVID-19 Decision Support System for Phone Call Triage, Designed by and for Medical Students, Stud Health Technol Inform. 281 (2021) 525–529. doi:10.3233/SHTI210226.

[5] Telion C, Marx JS, Dautreppe C, Carli P. Crise COVID-19 : les étudiants en médecine en renfort de la régulation médicale, Journal Européen Des Urgences et De Réanimation. 33 (2021) 96–101. doi:10.1016/j.jeurea.2021.05.001.

[6] Brasseur E, Servotte JC, Donneau AF, Stipulante S, d'Orio V, Ghuysen A., Triage for out-of-hours primary care calls: a reliability study of a new French-language algorithm, the SALOMON rule, Scand J Prim Health Care. 37 (2019) 227–232. doi:10.1080/02813432.2019.1608057.

[7] Cathain AO, Webber E, Nicholl J, Munro J, Knowles E. NHS Direct: consistency of triage outcomes, Emerg Med J. 20 (2003) 289–292. doi:10.1136/emj.20.3.289.

[8] North F, Richards DD, Bremseth KA, Lee MR, Cox DL, Varkey P, et al. Clinical decision support improves quality of telephone triage documentation - an analysis of triage documentation before and after computerized clinical decision support, BMC Med Inform Decis Mak. 14 (2014) 20. doi:10.1186/1472-6947-14-20.

[9] Dieckmann RA, Brownstein D, Gausche M. Hill, The pediatric assessment triangle: a novel approach for the rapid evaluation of children, Pediatr Emerg Care. 26 (2010) 312–315. doi:10.1097/pec.0b013e3181d6db37.

[10] Heude B, Scherdel P, Werner A, Guern ML, Gelbert N, Walther D, et al. A big-data approach to producing descriptive anthropometric references: a feasibility and validation study of paediatric growth charts, The Lancet Digital Health. 1 (2019) e413–e423. doi:10.1016/S2589-7500(19)30149-9.

[11] Madar R, Ugon A, Ivanković D, Tsopra R, A Web Interface for Antibiotic Prescription Recommendations in Primary Care: User-Centered Design Approach, J Med Internet Res. 23 (2021) e25741. doi:10.2196/25741.

Challenges of Trustable AI and Added-Value on Health
B. Séroussi et al. (Eds.)

doi:10.3233/SHTI220495

An Interactive Interface for Displaying Recommendations on Emergency Phone Triage in Pediatrics

Claire DURCHON[a,1*], Stephi VANDERLAN[a,1*], Alice JEGARD[a,1,*], Hasini SARAM[a,1*], Marina FALCHI[a], Florence CAMPEOTTO[a,b,c], Laurent DUPIC[b], Anita BURGUN [a,d,e,f], Benoît VIVIEN[a,b], and Rosy TSOPRA [a,d,e,f]
[a]*Digital Health Program of Université de Paris, Paris, France*
[b]*Régulation Régionale Pédiatrique, SAMU de Paris, Hôpital Necker - Enfants Malades, AP-HP, Paris, France*
[c]*Faculté de Pharmacie, Université de Paris, Inserm UMR S1139, Paris, France*
[d]*Université de Paris, Inserm, Sorbonne Université, Centre de Recherche des Cordeliers, F-75006 Paris, France*
[e]*HeKA, Inria Paris, France*
[f]*Department of Medical Informatics, Hôpital Européen Georges-Pompidou, AP-HP, Paris, France*
**Co-first authors, as they contributed equally to the work*

Abstract. Emergency phone triage aims at identifying quickly patients with critical emergencies. Patient triage is not an easy task, especially in situations involving children, mostly due to the lack of training and the lack of clinical guidelines for children. To overcome these issues, we aim at designing and assessing an interactive interface for displaying recommendations on emergency phone triage in pediatrics. Four medical students formalized local guidelines written by the SAMU of Paris, into a decision tree and designed an interface according to usability principles. The navigation within the interface was designed to allow the identification of critical emergencies at the beginning of the decision process, and thus ensuring a quick response in case of critical emergencies. The interface was assessed by 10 medical doctors: they appreciated the ergonomics (e.g., intuitive colors), and found easy to navigate through the interface. Nine of them would like to use this interface during phone call triage. In the future, this interface will be improved and implemented in emergency call centers.

Keywords. Clinical Decision Support System, Pediatric, Emergency, Triage, Usability, Digital Health.

1. Introduction

Emergency phone triage aims at identifying quickly patients with critical emergencies *versus* the others. They involve health professionals who assess patients remotely and

[1] Students from the Digital Health Program, Université de Paris; E-mail: claire.durchon@etu.u-paris.fr, stephi.vanderlan@etu.u-paris.fr, alice.jegard@etu.u-paris.fr, hasini.saram@etu.u-paris.fr

guide them to the appropriate levels of care: prehospital care, emergency department, GP at home or medical advice [1].

Patient triage is not an easy task, especially in situations involving children [2]. Indeed, in these situations, health professionals have to face to several issues [2]: (i) the difficulty to assess child remotely based on parents' speaking, (ii) the management of parents' anxiety, (iii) the lack of training and clinical practice guidelines in pediatrics.

To tackle these issues, a group of medical doctors from the SAMU of Paris (one of the biggest emergency call centers in France), develop recommendations for phone call triage of pediatric patients. They collaborated and validated together textual recommendations corresponding to a large number of medical situations that are frequently met in emergency pediatrics (e.g., acute dyspnea; head trauma, fever). However, these recommendations are entirely textual, complex and not easy to use in clinical practice, resulting in a risk of poor adoption by health professionals involved in phone triage [3]. Furthermore, they are not easy to share and not transportable. In the faculty of medicine, medical students involved in a digital health program, proposed to implement these recommendations into an interactive and intuitive interface.

Here, we aim at describing both the design and evaluation of this interface that display the recommendations for phone call triage of pediatric patients.

2. Methods

2.1 Interface design

Four medical students analyzed the textual recommendations written by the medical doctors of SAMU of Paris. For each medical situation, they manually extracted all the terms related to the decision making, made them unambiguous and categorized them into decision or action variables. Then, they organized them into a decision tree. The order of the decision variables into the decision tree, was decided to allow the identification of critical emergencies at the beginning of the decision process (e.g., "loss of consciousness" in head injury situation). This ensures a quick response in case of critical emergencies.

Design techniques were used to facilitate the navigation into the interface [4]: use of intuitive colors following the principle of traffic lights (e.g., red buttons for questions about severity symptoms, orange for additional questions, and green for the display of recommendations); use of information bubbles for helping health professionals to ask the correct questions to patients; display of questions in the form of tick boxes only when needed; organization of the interface into a structured way.

The interface was implemented using RStudio version 1.3.1093 and Shiny package. The design of the application was created with shinythemes, shinyjs, shinyBS, shinydashboard and shinyWidgets packages.

2.2 Interface evaluation

The interface was assessed online by 10 medical doctors with experience in emergency phone call triage or in the design of clinical decision support system. Some of them have been involved in the initial writing of the textual recommendations.

They tested the interface individually, without supervision, with two clinical cases (fever and head injury), and then filled an online form. The form collected anonymous sociodemographic data (e.g., level of experience in emergency phone call triage), and 6

questions related to the ergonomics and the utility of the interface compared to the textual recommendations. Each question was assessed through a 5-points Likert-Scale. Additional open-ended questions were also integrated for free comments.

3. Results

3.1 Interface

The interface is divided into two main areas (Figure 1). The first one contains information required for all clinical situations (age, reason of call). This area is displayed in the form of a column colored in black to attract the eye of the user.

The second area contains information related to the decision tree. This area is displayed in bigger and colored in white to allow better navigation into the decision tree. Four sub-areas can be distinguished:

- the blue area recaps the essential information in a concise way.
- the red area displays the first questions to ask to quickly assess the seriousness of the situation. These questions are adapted to the reason of call that has been selected in the first black area, and details can be provided on demand through information bubbles.
- the orange area displays additional questions, adapted to the reason of call. Questions are displayed progressively according to user responses.
- the green area displays the recommendation of the system. Five types of recommended actions can be displayed: “send an emergency ambulance”, “send the fire-fighters”, “send a GP at home”, “go to emergency by your own”, or “advice” (e.g., reassuring parents).

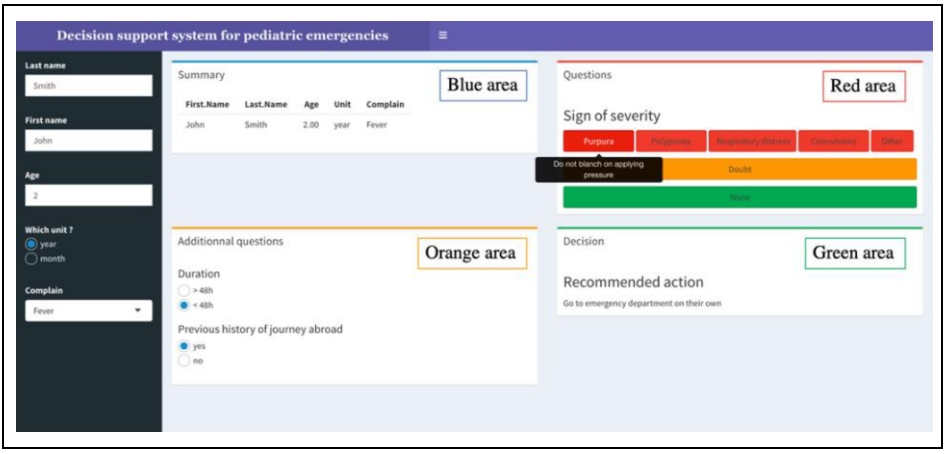


Figure 1. Interface. Example of a 2-year-old boy with fever. Severity signs to check in first are displayed within the red area (e.g., respiratory failure). If severity signs are absent, then additional questions are asked on the orange area (e.g., previous history of journey abroad). Finally, the recommended action is displayed in the green area (e.g., go to the emergency department). Additional information is available on interactive bubbles.

3.2 Evaluation

The interface was assessed by 4 senior doctors and 6 junior doctors/residents in emergency. Regarding experience in phone call triage: 7 were beginners, 2 intermediates, and 1 was confirmed.

90% of them would like to use this kind of interface during phone call triage (Figure 2). All of them liked the interface design and found the colors intuitive. 80% of them found easy to navigate through the interface. The qualitative analysis of open questions highlighted that the interface was easy and quick to use. One doctor proposed to consider other parameters in the medical decision, such as the tone of voice for patient anxiety, especially in case of doubt. Another explained that it could be hard to use this interface during the phone call triage, because of the difficulty of patient interview in real life.

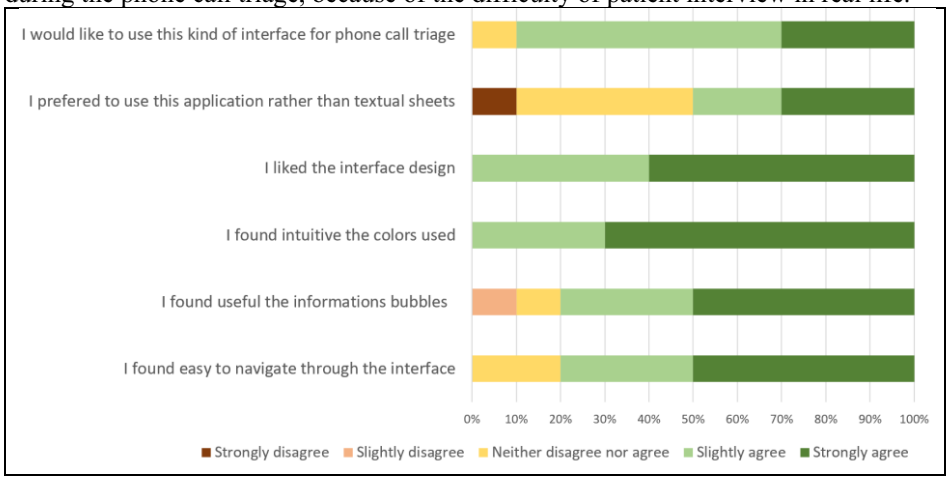


Figure 2. Interface evaluation.

4. Discussion

We designed an interface that displayed the recommendations for phone call triage of pediatric patients. Recommendations were issued from textual recommendations designed by medical experts from SAMU of Paris, one of the biggest emergency call centers in France. Recommendations were formalized into a decision tree by four medical students, and then implemented into an interface following usability principles and design techniques. Ten medical doctors assessed the interface: 90% of them would like to use it during phone call triage, and all of them liked interface design. The interface seems to be sufficiently easy and intuitive to use, without the need of training or tutorial.

Some limits need to be highlighted. First, the update of the interface can cause problem since the medical knowledge is constantly changing. To overcome this issue, we plan to develop an administrator interface allowing medical experts to automatically update the interface, without the need of a computer engineer. Another possibility would be to use Natural Language Processing techniques [5] to automatically build the decision trees from textual recommendations. Second, the interface considers only a few clinical situations met in phone call triage. In the future we aim at implementing all the most frequent clinical situations met in emergency call centers and validating the related

decision trees by medical doctors. Third, the interface was assessed by only 10 medical doctors with only two clinical cases, and the evaluation focused on ergonomics and the utility of the interface. However, this work is the first step of software lifecycle, and these preliminary results will help us to improve the current interface and then conduct a larger trial. In the future we plan to consider other parameters mentioned by the evaluators, such as more help details in case of doubt [6] or for managing patient anxiety. Other parameters will also be considered, such as those having showed their utility in decision support system efficacy (e.g., the provision of advice for both patients and doctors) [7].

Finally, this work was part of a new digital health program at Université de Paris. This innovative program aims at involving both medical students and doctors in the design of clinical decision support system (under the supervision of the medical informatic team). This work demonstrated the need to develop such digital health program in medical curriculum, and not only at master's degree level [8,9].

Funding

This work is part of an innovative program named "Artificial Intelligence for Clinical Decision Support", dedicated to undergraduate medical students, at Université de Paris.

References

[1] Thierrin C, Augsburger A, Dami F, Monney C, Staeger P, Clair C. Impact of a telephone triage service for non-critical emergencies in Switzerland: A cross-sectional study, PLoS One. 16 (2021) e0249287. doi:10.1371/journal.pone.0249287.

[2] Koziel JR, Meckler G, Brown L, Acker D, Torino M, Walsh B, et al. Barriers to pediatric disaster triage: a qualitative investigation, Prehosp Emerg Care. 19 (2015) 279–286. doi:10.3109/10903127.2014.967428.

[3] Cabana MD, Rand CS, Powe NR, Wu AW, Wilson MH, Abboud PA et al. Why don't physicians follow clinical practice guidelines? A framework for improvement, JAMA. 282 (1999) 1458–1465.

[4] Miller K, Mosby D, Capan M, Kowalski R, Ratwani R , Noaiseh Y et al. Interface, information, interaction: a narrative review of design and functional requirements for clinical decision support, J Am Med Inform Assoc. 25 (2017) 585–592. doi:10.1093/jamia/ocx118.

[5] Nadkarni PM, Ohno-Machado L, Chapman WW. Natural language processing: an introduction, J Am Med Inform Assoc. 18 (2011) 544–551. doi:10.1136/amiajnl-2011-000464.

[6] Tsopra R, Sedki K, Courtine M, Falcoff H, De Beco A, Madar R, et al. Helping GPs to extrapolate guideline recommendations to patients for whom there are no explicit recommendations, through the visualization of drug properties. The example of AntibioHelp® in bacterial diseases, J Am Med Inform Assoc. 26 (2019) 1010–1019. doi:10.1093/jamia/ocz057.

[7] Roshanov PS, Fernandes N, Wilczynski JM, Hemens BJ, You JJ, Handler SM, et al., Features of effective computerised clinical decision support systems: meta-regression of 162 randomised trials, BMJ. 346 (2013) f657. doi:10.1136/bmj.f657.

[8] Ammenwerth E, First Accreditation of a Medical Informatics Program by the European Federation of Medical Informatics: Experiences, Stud Health Technol Inform. 279 (2021) 34–35. doi:10.3233/SHTI210085.

[9] Valenta AL, Berner ES, Boren SA, Deckard DJ, Eldredge C, Fridsma DB, et al. AMIA Board White Paper: AMIA 2017 core competencies for applied health informatics education at the master's degree level, J Am Med Inform Assoc. 25 (2018) 1657–1668. doi:10.1093/jamia/ocy132.

Challenges of Trustable AI and Added-Value on Health
B. Séroussi et al. (Eds.)

doi:10.3233/SHTI220496

Optimization of Performance by Combining Most Sensitive and Specific Models in Data Science Results in Majority Voting Ensemble

Katoo M. MUYLLE[a], Pieter CORNU[a], Wilfried COOLS[b], Kurt BARBÉ[b], Ronald BUYL[b] and Sven VAN LAERE[b,1]
[a] *Centre for Pharmaceutical Research (CePhar), Vrije Universiteit Brussel, Belgium*
[b] *Department of Public Health, Vrije Universiteit Brussel, Belgium*

Abstract. Ensemble modeling is an increasingly popular data science technique that combines the knowledge of multiple base learners to enhance predictive performance. In this paper, the idea was to increase predictive performance by holding out three algorithms when testing multiple classifiers: (a) the best overall performing algorithm (based on the harmonic mean of sensitivity and specificity (HMSS) of that algorithm); (b) the most sensitive model; and (c) the most specific model. This approach boils down to majority voting between the predictions of these three base learners. In this exemplary study, a case of identifying a prolonged QT interval after administering a drug-drug interaction with increased risk of QT prolongation (QT-DDI) is presented. Performance measures included accuracy, sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV). Overall performance was measured by calculating the HMSS. Results show an increase in all performance measure characteristics compared to the original best performing algorithm, except for specificity where performance remained stable. The presented approach is fairly simple and shows potential to increase predictive performance, even without adjusting the default cut-offs to differentiate between high and low risk cases. Future research should look at a way of combining all tested algorithms, instead of using only three. Similarly, this approach should be tested on a multiclass prediction problem.

Keywords. Artificial Intelligence, Performance Measures, Majority Voting, Ensemble Learning, Drug Interactions

1. Introduction

In the field of data science, it is common practice to use the ensemble modeling technique. In that technique, the word *ensemble* literally means together, where the data scientist uses different base models that work together to get a final outcome.[1] In ensemble modeling, literature often refers to two basic techniques, i.e. bagging and boosting.[2-4]

In bagging, the researcher first selects a random sample that serves as input to a certain model to train that model, and repeats this process multiple times. In the end, all different trained models are aggregated to reduce the variance, for example by using the majority voting technique.[4, 5] In boosting, different models are build sequentially,

[1] Sven Van Laere (Corresponding author), Department of Public Health, Vrije Universiteit Brussel, Laarbeeklaan 103, 1090 Jette, BELGIUM; E-mail: Sven.Van.Laere@vub.be.

where the output of the previous model is used as input for the next model in order to reduce bias in the outcome.[6] Eventually, all models are then used together to obtain an overall prediction.

In this manuscript, the authors did not prespecify what is meant by 'models' in the explanation of bagging and boosting written above. The simplest forms of these models are a decision tree (in case of a categorical outcome) or a decision tree regressor (in case of a continuous outcome), but other modelling techniques (e.g. linear or logistic regression models) can also be used in these ensemble modelling techniques.

The performance of a final model is often expressed in terms of accuracy, precision, recall, or another performance characteristic.[7] In general, it is difficult to decide which model is the best model on overall, since different performance characteristics tend to focus on different desirable characteristics (e.g. having few false negatives). Therefore, researchers invented performance measures like the F1-score that combines both the performance measures precision (positive predictive value, PPV) and recall (sensitivity) by taking its harmonic mean to be able to compare models.

In a clinical setting, such models can be incorporated into clinical decision support systems (CDSS) to determine when an alert to a healthcare professional should be triggered to support decision-making. The developer of such a CDSS is not solely interested in identifying positive cases (i.e., patients with a certain disease or symptom), but should also focus on negative cases (i.e., patients that do not have a certain disease or symptom).[8] This can be explained by the fact that a major drawback of improving the sensitivity of a CDSS is that it becomes less specific, potentially leading to alert fatigue of that healthcare professional.[9]

In this work, the authors are interested in increasing overall performance by complementing the best performing model (selected based on its harmonic mean of sensitivity and specificity (HMSS)) with models having the highest sensitivity and the highest specificity.

2. Methods

In an effort to increase model performance on a dataset for predicting QT prolongation after exposure to a drug-drug interaction with increased risk of QT prolongation (QT-DDI), the authors used several algorithms (logistic regression, Gaussian Naive Bayes classifier, support vector machine, random forest, gradient boosting, etc.) to predict high and low risk patients. Prevention of drug-induced QT prolongation can in turn prevent a potentially lethal type of ventricular tachycardia called Torsades de Pointes (TdP)). In a systematic review, Arunachalam et al.[10] reported an incidence of drug-induced QT prolongation of 6.3%, while the incidence of TdP was 0.33% in patients exposed to drugs that prolong the QT interval.

In an internal validation phase, 350 patients were modelled using 10-fold stratified cross-validation. Subsequently, an external validation was performed on 110 new patient cases to find a best overall model. Performance measures were reported in terms of accuracy, sensitivity, specificity, PPV, and negative predictive value (NPV). Overall performance was measured in terms of HMSS. The authors combined the best overall performing algorithm, with the most sensitive and the most specific model as demonstrated in Table 1.

This new approach of combining prediction results can be viewed as using one default prediction outcome (algorithm 1), with two complementary prediction outcomes

of which one focusses on predicting positive cases (i.e., high risk cases – algorithm 2) and one focusses on predicting negative cases (i.e., low risk cases – algorithm 3). When algorithm 2 predicts a high risk case, the combined prediction will predict a high risk case as outcome. Only when algorithm 3 predicts a low risk case at the same time, the combined outcome falls back to the default algorithm 1. Similarly, when algorithm 3 predicts a low risk case, the combined algorithm will return this prediction. Also here, the combined algorithm falls back to the default algorithm 1 when algorithm 2 predicts to be a high risk case and algorithm 3 predicts to be a low risk case. In all other cases, the default prediction outcome of the best overall algorithm 1 was used.

Table 1. Prediction strategy combining the best overall, sensitive (precision) and specific (recall) algorithm.

	Prediction algorithm 1 *(best overall)*	*Prediction algorithm 2* *(best sensitivity)*	*Prediction algorithm 3* *(best specificity)*	*Combined prediction*	*Algorithm to follow*
Case 1	Low risk	Low risk	Low risk	Low risk	3
Case 2	Low risk	Low risk	High risk	Low risk	1
Case 3	Low risk	High risk	Low risk	Low risk	1
Case 4	Low risk	High risk	High risk	High risk	2
Case 5	High risk	Low risk	Low risk	Low risk	3
Case 6	High risk	Low risk	High risk	High risk	1
Case 7	High risk	High risk	Low risk	High risk	1
Case 8	High risk	High risk	High risk	High risk	2

3. Results

In Table 2, the results of the external validation were reported. The performance increased by combining the best overall performing algorithm (based on HMSS) with the best sensitive and the best specific algorithm.

The overall performance, measured by the HMSS, of that combined algorithm increased with about 3% (i.e., from 63.49% in the best overall performing model (algorithm 1) to 66.57% in the new combined form). The accuracy increased with about 5.5% (from 63.64% to 69.09%), the sensitivity increased with about 6% (from 62.50% to 68.75%), the specificity remained the same (64.52%), the PPV increased with about 2.5% (from 57.69% to 60.00%) and the NPV increased with about 4% (from 68.97% to 72.73%). (see Table 2)

This combined form of prediction can – in a way – be interpreted as a special form of majority voting ensemble. The *ensemble* part in that wording refers to the three algorithms that are used to construct a new combined prediction and the *majority voting* part in that wording refers to the combined prediction that is in fact based on the majority of votes over the three algorithms (see Table 1). The majority vote resulted in a positive or negative case (i.e. respectively high or low risk for QT prolongation after administering a QT-DDI) when respectively two or three of these selected models predicted a positive or negative case.

Table 2. Summary of performance characteristics obtained after external validation. Performance characteristics increased by applying ensemble modeling of the best performing algorithms.

Algorithm number	1	2	3	4	…	Combined
Accuracy	0.6364	0.4818	0.6909	0.6182		0.6909
Sensitivity	0.6250	**0.9375**	0.4167	0.5833		0.6875

Specificity	0.6452	0.1290	**0.9032**	0.6452		0.6452
PPV	0.5769	0.4545	0.7692	0.5600		0.6000
NPV	0.6897	0.7273	0.6667	0.6667		0.7273
Harmonic mean	***0.6349***	*0.2268*	*0.5703*	*0.6127*		*0.6657*

Algorithm 1: Logistic regression with forward model building (best overall); Algorithm2: Gaussian Naive Bayes classifier (best sensitivity); Algorithm 3: Random forest with feature selection (best specificity); Algorithm 4: Random forest with feature selection; PPV: positive predictive value; NPV: negative predictive value; Harmonic mean: harmonic mean of sensitivity and specificity (HMSS). Only algorithm 1, 2 and 3 were used to construct the new combined prediction outcome.

4. Discussion

This study presents a way of increasing predictive performance by applying majority voting ensemble by combining the three most performant algorithms in terms of overall performance, sensitivity and specificity. The reason for using sensitivity and specificity stems from the fact that CDSS should both focus on identifying the positive and negative cases in order to keep alert fatigue as low as possible. This paper presents a case of finding an optimized prediction of a high risk for QT prolongation in patients exposed to QT-DDIs. It was shown that, by combining these three models or algorithms, all performance characteristics increased or remained equal to that of the original best performing prediction model.

The strength of this approach is that it is fairly simple to obtain this increased performance: (a) consider a number of prediction algorithms, (b) conduct the training and evaluation of these models based on the internal validation dataset, (c) test all algorithms with an external validation dataset, and then (d) constitute a new combined model based on the performance characteristics of all models. A limitation of this approach is that it requires many different models before being able to select a model with a high enough sensitivity and a high enough specificity in order to find a possibility to increase overall performance. Another limitation to this approach is that it depends on only two measures (i.e., sensitivity and specificity). Moreover, these two performance measures can be increased or decreased according to the chosen cut-off that the researcher applies. For example, in logistic regression, a traditional cut-off is 0.50 for the predicted probability of a certain event (e.g., QT prolongation) in order to predict whether an event will occur or not.[11] Sensitivity analysis might find a more optimal cut-off in order to optimize the sensitivity-specificity trade-off. Another limitation is that algorithms that have 100% sensitivity or 100% specificity, e.g. the case when the algorithm either classifies each data instance as positive or negative, will not increase the performance.

Therefore, future work will additionally implement a sensitivity analysis on all algorithms when performing the external validation in order to find an appropriate cut-off before combining classifiers. This will prevent algorithms from having an utterly high/low sensitivity or an utterly high/low specificity. In this paper, we showed an example of a binary outcome (high or low risk for QT prolongation) by testing the performance of different classifiers. Future work should also assess the potential of using this technique in a multiclass setting. Moreover, to prevent this new approach from only taking the three best performing algorithms (respectively in terms of overall performance, sensitivity and specificity), future work might exist in finding a way to combine the information of all different models to boost the overall performance in an optimal way.

5. Conclusion

In this study, the authors showed that combining the best performing overall classifier with the classifiers having the best sensitivity and best specificity, increased predictive performance. The approach is fairly simple once having multiple trained classifiers ready to be tested on an external dataset. The presented method boils down to a type of majority voting ensemble where three algorithms are combined using the most present predicted outcome as final outcome.

References

[1] Zhou ZH, "Ensemble Learning," in Machine Learning, Zhou ZH, Ed. Singapore: Springer Singapore, 2021, pp. 181-210. doi: https://doi.org/10.1007/978-981-15-1967-3_8.
[2] Breiman L, "Bagging predictors," Machine Learning, vol. 24, no. 2, pp. 123-140, 1996/08/01 1996, doi: https://doi.org/10.1007/BF00058655.
[3] Freund Y, Schapire RE. "A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting," Journal of Computer and System Sciences, vol. 55, no. 1, pp. 119-139, 1997/08/01/ 1997, doi: https://doi.org/10.1006/jcss.1997.1504.
[4] Kuncheva LI, Skurichina M, Duin RPW. "An experimental study on diversity for bagging and boosting with linear classifiers," Information Fusion, vol. 3, no. 4, pp. 245-258, 2002/12/01/ 2002, doi: https://doi.org/10.1016/S1566-2535(02)00093-3.
[5] Altman N, Krzywinski M. "Ensemble methods: bagging and random forests," Nature Methods, vol. 14, no. 10, pp. 933-934, 2017/10/01 2017, doi: https://doi.org/10.1038/nmeth.4438.
[6] Webb GI, Zheng Z. "Multistrategy ensemble learning: reducing error by combining ensemble learning techniques," IEEE Transactions on Knowledge and Data Engineering, vol. 16, no. 8, pp. 980-991, 2004, doi: https://doi.org/10.1109/TKDE.2004.29.
[7] Sokolova M, Japkowicz N, Szpakowicz S. "Beyond Accuracy, F-Score and ROC: A Family of Discriminant Measures for Performance Evaluation," in AI 2006: Advances in Artificial Intelligence, Berlin, Heidelberg, A. Sattar and B.-h. Kang, Eds., 2006// 2006: Springer Berlin Heidelberg, pp. 1015-1021. doi: https://doi.org/10.1007/11941439_114.
[8] Payne TH, et al. "Recommendations to improve the usability of drug-drug interaction clinical decision support alerts," (in eng), J Am Med Inform Assoc, vol. 22, no. 6, pp. 1243-50, Nov 2015, doi: https://doi.org/10.1093/jamia/ocv011.
[9] Légat L, Van Laere S, Nyssen M, Steurbaut S, Dupont AG, Cornu P. "Clinical Decision Support Systems for Drug Allergy Checking: Systematic Review," (in eng), Journal of medical Internet research, vol. 20, no. 9, pp. e258-e258, 2018, doi: https://doi.org/10.2196/jmir.8206.
[10] Arunachalam K, Lakshmanan S, Maan A, Kumar N, Dominic P. "Impact of Drug Induced Long QT Syndrome: A Systematic Review," (in eng), J Clin Med Res, vol. 10, no. 5, pp. 384-390, May 2018, doi: https://doi.org/10.14740/jocmr3338w.
[11] Hosmer DW, Lemeshow S, Sturdivant RX. Applied Logistic Regression, 3rd Edition ed. Wiley, 2013.

Challenges of Trustable AI and Added-Value on Health
B. Séroussi et al. (Eds.)

doi:10.3233/SHTI220497

Implementation of a Clinical Trial Recruitment Support System Based on Fast Healthcare Interoperability Resources (FHIR) in a Cardiology Department

Clemens SCHERER [a,b,1], Stephan ENDRES [c], Martin ORBAN [a,b], Stefan KÄÄB [a,b], Steffen MASSBERG [a,b], Alfred WINTER [d] and Matthias LÖBE [d]
[a] *Department of Medicine I, University Hospital, LMU Munich, Munich, Germany*
[b] *DZHK (German Center for Cardiovascular Research), Partner Site Munich Heart Alliance, Department of Medicine I, University Hospital, LMU Munich*
[c] *Medical Technology and IT (MIT), University Hospital, LMU Munich, Munich, Germany*
[d] *Institute for Medical Informatics, Statistics and Epidemiology (IMISE), University of Leipzig, Leipzig, Germany*

Abstract. Clinical Trial Recruitment Support Systems can booster patient inclusion of clinical trials by automatically analyzing eligibility criteria based on electronic health records. However, missing interoperability has hindered introduction of those systems on a broader scale. Therefore, our aim was to develop a recruitment support system based on FHIR R4 and evaluate its usage and features in a cardiology department. Clinical conditions, anamnesis, examinations, allergies, medication, laboratory data and echocardiography results were imported as FHIR resources. Clinical trial information, eligibility criteria and recruitment status were recorded using the appropriate FHIR resources without extensions. Eligibility criteria linked by the logical operation "OR" were represented by using multiple FHIR Group resources for enrollment. The system was able to identify 52 of 55 patients included in four clinical trials. In conclusion, use of FHIR for defining eligibility criteria of clinical trials may facilitate interoperability and allow automatic screening for eligible patients at multiple sites of different healthcare providers in the future. Upcoming changes in FHIR should allow easier description of "OR"-linked eligibility criteria.

Keywords. *Clinical trial recruitment support system, FHIR, cardiology, research study, eligibility criteria, interoperability*

1. Introduction

Clinical trials are the mainstay of medical advancement in modern times enabling so-called evidence-based medicine [1]. However, tenacious patient recruitment hampers medical advancement [2]. In many studies, the actual number of recruitable patients is overrated in the planning phase [3]. This misconception leads to increasing trial costs

[1] Corresponding Author: Dr. med. Clemens Scherer, Department of Medicine I, University Hospital, LMU Munich, Marchioninistr. 15, 81377 Munich, Germany; E-mail: clemens.scherer@med.uni-muenchen.de.

and possibly discontinuation of a trial [4]. Insufficient recruiting of patients may be due to an actual overestimation of patients fulfilling trial criteria, insufficient identification of those patients by a healthcare provider or lack of will to participate [5].

In the past decade, support systems have been developed to assess inclusion and exclusion criteria of a trial based on patients' electronic healthcare records (EHR). More specifically, clinical trial recruitment support systems (CTRSS), which retrospectively query existing data in EHR, can be distinguished from clinical trial alert systems, which monitor the occurrence of a specific event and generate an alert [6]. Ideally, generation of false notifications should be avoided to preserve physicians' participation and accurate notification should come in a suitable time window in which the patient can be approached [7].

Healthcare system-wide introduction of CTRSS is still pending in many countries. In the beginning, attention has been directed to technical feasibility neglecting clinical practicability [8]. Also, the complete and accurate mapping of the inclusion and exclusion criteria and their translation into structured queryable variables is complex. In addition, these variables must be present in the EHR system in a structured manner and must be documented for specific patients, whereby phenotyping techniques can be applied [9]. Furthermore, missing interoperability has hindered transfer to other applications [10].

Originally introduced in 2011, Fast Healthcare Interoperability Resources (FHIR) may become the new standard for interoperability in healthcare enabling communication of many areas of the healthcare system. Popular terminology systems used in conjunction with FHIR are “Systematized Nomenclature of Medicine – Clinical Terms” (SNOMED CT) and “Logical Observation Identifiers Names and Codes” (LOINC) [11, 12]. Clinical trials and eligibility criteria are defined in FHIR by the resources “ResearchStudy” and “Group”. However, both have a maturity level of 1 indicating “trial use” with expected major changes in the future. A storage of clinical trial records using FHIR was already implemented in a recent study, however without emphasis on enrollment criteria [13]. The aim of our study was to develop a CTRSS based on FHIR R4 and to evaluate its usage and features in a cardiology department based on clinical trials conducted there.

2. Methods

Patient data on medical conditions, anamnesis, examination, medication and allergies were already available on a self-developed FHIR server and could be accessed via the FHIR resources “Condition”, “AllergyIntolerance”, “Observation” and “MedicationStatement”. Physicians were required to input those in the first 24 hours after admission. Observations were coded using the terminology systems SNOMED CT and LOINC. However, only a fraction of medical conditions including the most abundant cardiologic conditions was coded by SNOMED CT due to the limited time of physicians to add a code to all conditions during patients stay (diagnoses in Germany are normatively coded with ICD-10-GM, mainly after discharge). All other conditions were recorded as text and were classified by simple text pattern recognition. Medication was coded by their Anatomical Therapeutic Chemical Classification System (ATC) codes.

Relevant laboratory data was imported from the hospital's electronic health record, mapped to FHIR with a LOINC terminology mapping table using Unified Code for Units of Measure (UCUM) codes and saved in the local FHIR server of the cardiology department. Furthermore, trial-relevant echocardiographic measurements were imported

from a Microsoft SQL database of the echocardiography system. However, only around 50% of patients had the left ventricular ejection fraction, which is an important parameter for trials in cardiology, numerically recorded. In all other cases, ejection fraction was solely recorded as text in the report. Therefore, ejection fraction was extracted by simple text pattern recognition and saved as an "Observation" on the department's local FHIR server. Both laboratory and echocardiographic data were updated every 10 minutes from the source database.

A clinical trial was defined by the FHIR resource "ResearchStudy". Eligibility criteria were recorded using the element "eligibility" of "ResearchStudy", which itself is an array of the generic FHIR resource "Group". Groups are collections that either have explicit members or are defined by a set of characteristics. In such a "Group", a set of eligibility criteria can be defined by the Backbone element array "characteristic". The code of a criteria is defined by "code" and the value by "value[x]". Additionally, a characteristic element can be negated by using the Boolean element "exclude" allowing representation of an exclusion criteria. All "characteristics" are linked by the logical operator "AND" by definition. However, some trials link their eligibility criteria by the logical operator "OR". In FHIR R4, this connection can only be represented by using multiple groups without introduction of extensions. For example, having two independent "OR" connections with two eligibility criteria each would results in four FHIR "Group" resources for the element "enrollment".

Trial study nurses and physicians were enabled to add new and edit trial information and input inclusion and exclusion criteria using a web-browser user interface in the hospital intranet. All information were recorded on the server side as the FHIR resources "ResearchStudy" and "Group". On the client side, eligibility criteria were transformed to a tree-like structure (**Figure 1**). Upon user demand, all hospitalized and ambulatory patients in the cardiology department were instantly screened for trial eligibility using the FHIR eligibility criteria on the existing patients' FHIR resources. Furthermore, study personal was able to manually edit trial status (i.e. ineligible, on-study, ...) of patients, which was implemented using the FHIR resource "ResearchSubject".

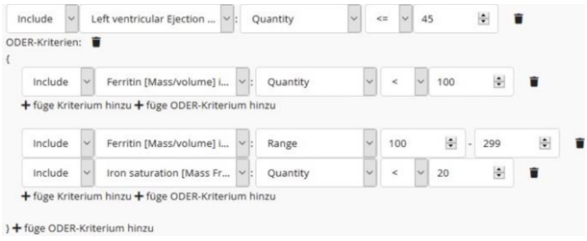


Figure 1. Screenshot of the interface to construct inclusion/exclusion criteria groups.

3. Results

This implementation of a CTRSS based on FHIR R4 was evaluated in clinical practice: Beginning from 1st April 2021 the application was used as an additional patient screening tool for the four trials CLOSURE-AF, FAIR-HF2, SPRIRIT-HF and TORCH-PLUS of the German Centre for Cardiovascular Research (DZHK) at the site "Großhadern" of the Department of Cardiology of LMU University Hospital. Those four trials were selected because of the known difficulty to identify patients for those trials.

As the COVID-19 pandemic is prohibiting any proper comparison of patient inclusion rates, efficacy of the recruitment support system was tested by comparing the numbers of patients identified by the recruitment support system and enrolled in a trial to the actual number of enrolled patients irrespective of the screening method from 1st April 2021 to 23rd November 2021. 52 out of 55 patients could be identified by the system (see **Table 1**). Missing user input and rapid changes in laboratory values were the main reasons for the failed screening.

Table 1. Efficacy of identification of eligible patients

Clinical trial name	Total number of patients enrolled	Number of enrolled patients identified by the recruitment support system
CLOSURE-AF	5	4
FAIR-HF	4	4
SPIRIT-HF	2	2
TORCH-PLUS	44	42

4. Discussion

This implementation of a CTRSS could show that interoperability regarding trial eligibility criteria can be achieved by using FHIR R4 resources, even for complex criteria interactions using the logical operator "OR". For better screening performance, eligibility criteria were altered for screening: For example, laboratory values may change over the course of a hospital stay and a wider range should be applied to identify patients. Updated FHIR representations of "ResearchStudy" should distinguish between eligibility and screening criteria to overcome this issue. Moreover, a relative time period element dating back from the screening date should be included in the FHIR resource "Group" to allow implementation of eligibility criteria such as "no hemoglobin drop < 9 g/dL during the last 30 days".

In this implementation, patients who did not fulfill all eligibility criteria due to missing values were included in the screening list as "possible candidates". Those missing values could be due incomplete user input or the necessity of further examinations, which would require patient consent. A more in-depth dissection of unknown values may facilitate user interaction in the future.

FHIR resources encompassing clinical trials and eligibility may be prone to changes in the future due to their maturity status which may facilitate implementations: For example, allowing the backbone members in the FHIR resource Group for descriptive groups may allow a more straightforward way to describe a set of eligibility criteria which are connected by the logical operation "OR".

5. Conclusions

In this project, the authors developed a CTRSS based on FHIR R4, which was able to detect suitable patients based on eligibility criteria in an efficient way. Therefore, the use of FHIR for defining eligibility criteria of clinical trials may facilitate interoperability and allow automatic screening for eligible patients at multiple sites of different healthcare providers, thereby optimizing recruitment of patients for clinical trials. This

is particularly important in large national initiatives currently underway, such as the Medical Informatics Initiative (MII) [14], which aims to make the healthcare data of all university hospitals in Germany available for clinical research, or the National Research Data Infrastructure (NFDI4Health) [15], which provides clinical, epidemiological and public health studies in the course of data sharing.

Acknowledgments

The work is supported by the corresponding grants (BMBF 01ZZ1803A, DFG 442326535, DFG 413635475).

References

[1] Backmann M. What's in a gold standard? In defence of randomised controlled trials. Med Health Care Philos. 2017;20(4):513-23. Epub 2017/04/23. doi: 10.1007/s11019-017-9773-2.
[2] Nkoy FL, Wolfe D, Hales JW, Lattin G, Rackham M, Maloney CG. Enhancing an existing clinical information system to improve study recruitment and census gathering efficiency. AMIA Annu Symp Proc. 2009;2009:476-80. Epub 2009/01/01. PubMed PMID: 20351902.
[3] McDonald AM, Knight RC, Campbell MK, Entwistle VA, Grant AM, Cook JA, et al. What influences recruitment to randomised controlled trials? A review of trials funded by two UK funding agencies. Trials. 2006;7:9. Epub 2006/04/11. doi: 10.1186/1745-6215-7-9. PubMed PMID: 16603070.
[4] Collins JF, Williford WO, Weiss DG, Bingham SF, Klett CJ. Planning patient recruitment: fantasy and reality. Stat Med. 1984;3(4):435-43. Epub 1984/10/01. doi: 10.1002/sim.4780030425.
[5] Ohmann C, Kuchinke W. Meeting the Challenges of Patient Recruitment. International Journal of Pharmaceutical Medicine. 2007;21(4):263-70. doi: 10.2165/00124363-200721040-00002.
[6] Becker L, Gansland T, Prokosch H-U, Newe A. Recruitment for Clinical Trials – A Challenge for Electronic Support? Research Square. 2021. doi: 10.21203/rs.2.385/v1.
[7] Embi PJ, Jain A, Clark J, Harris CM. Development of an electronic health record-based Clinical Trial Alert system to enhance recruitment at the point of care. AMIA Annu Symp Proc. 2005;2005:231-5. Epub 2006/06/17. PubMed PMID: 16779036; PubMed Central PMCID: PMCPMC1560758.
[8] Cuggia M, Besana P, Glasspool D. Comparing semi-automatic systems for recruitment of patients to clinical trials. Int J Med Inform. 2011;80(6):371-88. Epub 2011/04/05. doi: 10.1016/j.ijmedinf.2011.02.003. PubMed PMID: 21459664.
[9] Löbe M, Stäubert S, Goldberg C, Haffner I, Winter A. Towards Phenotyping of Clinical Trial Eligibility Criteria. Stud Health Technol Inform. 2018;248:293-9. Epub 2018/05/05. PubMed PMID: 29726450.
[10] Köpcke F, Prokosch HU. Employing computers for the recruitment into clinical trials: a comprehensive systematic review. J Med Internet Res. 2014;16(7):e161. Epub 2014/07/06. doi: 10.2196/jmir.3446. PubMed PMID: 24985568; PubMed Central PMCID: PMCPMC4128959.
[11] Lee D, Cornet R, Lau F, de Keizer N. A survey of SNOMED CT implementations. Journal of Biomedical Informatics. 2013;46(1):87-96. doi: https://doi.org/10.1016/j.jbi.2012.09.006.
[12] McDonald CJ, Huff SM, Suico JG, Hill G, Leavelle D, Aller R, et al. LOINC, a Universal Standard for Identifying Laboratory Observations: A 5-Year Update. Clinical Chemistry. 2003;49(4):624-33. doi: 10.1373/49.4.624.
[13] Gulden C, Blasini R, Nassirian A, Stein A, Altun FB, Kirchner M, et al. Prototypical Clinical Trial Registry Based on Fast Healthcare Interoperability Resources (FHIR): Design and Implementation Study. JMIR Med Inform. 2021;9(1):e20470. Epub 2021/01/13. doi: 10.2196/20470. PubMed PMID: 33433393; PubMed Central PMCID: PMCPMC7837997.
[14] Semler SC, Wissing F, Heyder R. German Medical Informatics Initiative. Methods Inf Med. 2018;57(S 01):e50-e6. Epub 2018/07/18. doi: 10.3414/me18-03-0003. PubMed PMID: 30016818; PubMed Central PMCID: PMCPMC6178199.
[15] Schmidt CO, Darms J, Shutsko A, Löbe M, Nagrani R, Seifert B, et al. Facilitating Study and Item Level Browsing for Clinical and Epidemiological COVID-19 Studies. Stud Health Technol Inform. 2021;281:794-8. Epub 2021/05/28. doi: 10.3233/shti210284. PubMed PMID: 34042687.

Challenges of Trustable AI and Added-Value on Health
B. Séroussi et al. (Eds.)

doi:10.3233/SHTI220498

Out-of-Hospital Cardiac Arrest Detection by Machine Learning Based on the Phonetic Characteristics of the Caller's Voice

Sonia RAFI[a]; Cedric GANGLOFF[a]; Etienne PAULHET[b]; Ollivier GRIMAULT[cd]; Louis SOULAT[b]; Guillaume BOUZILLÉ[a]; Marc CUGGIA[a]
[a] *Univ Rennes, CHU Rennes, INSERM, LTSI-UMR 1099, F-35000 Rennes, France*
[b] *Department of Emergency Medicine, CHU Pontchaillou, Rennes, France*
[c] *Search And Rescue team, Department of Emergency Medicine, CHU de la Cavale Blanche, Brest, France*
[d] *EA4324-ORPHY, Université de Bretagne Occidentale, UFR Sciences et Techniques, Brest, France*

Abstract. Introduction. Out-of-hospital cardiac arrest (OHCA) is a major public health issue. The prognosis is closely related to the time from collapse to return of spontaneous circulation. Resuscitation efforts are frequently initiated at the request of emergency call center professionals who are specifically trained to identify critical conditions over the phone. However, 25% of OHCAs are not recognized during the first call. Therefore, it would be interesting to develop automated computer systems to recognize OHCA on the phone. The aim of this study was to build and evaluate machine learning models for OHCA recognition based on the phonetic characteristics of the caller's voice. **Methods.** All patients for whom a call was done to the emergency call center of Rennes, France, between 01/01/2017 and 01/01/2019 were eligible. The predicted variable was OHCA presence. Predicting variables were collected by computer-automatized phonetic analysis of the call. They were based on the following voice parameters: fundamental frequency, formants, intensity, jitter, shimmer, harmonic to noise ratio, number of voice breaks, and number of periods. Three models were generated using binary logistic regression, random forest, and neural network. The area under the curve (AUC) was the primary outcome used to evaluate each model performance. **Results.** 820 patients were included in the study. The best model to predict OHCA was random forest (AUC=74.9, 95% CI=67.4-82.4). **Conclusion.** Machine learning models based on the acoustic characteristics of the caller's voice can recognize OHCA. The integration of the acoustic parameters identified in this study will help to design decision-making support systems to improve OHCA detection over the phone.

Keywords. cardiac arrest, resuscitation, dispatcher, call center, acoustic, phonetic, voice analysis, artificial intelligence, machine learning.

1. Introduction

Out-of-hospital cardiac arrest (OHCA) is a major public health concern [1]. The prognosis is closely related to the time from collapse to return of spontaneous circulation [2]. The resuscitation efforts are frequently initiated at the request of emergency call center professionals who are specifically trained to identify critical conditions over the phone. However, 25% of OHCAs are not recognized during the first call, most often because

emergency call centers are overwhelmed [3]. In this context, it would be interesting to develop automated computer systems to recognize OHCA based on the bystanders' speech on the phone.

Speech analysis can be decomposed in two fields: linguistic and phonetic. Linguistic analysis investigates the meaning of words and their relationships, while phonetic analysis focuses on the voice acoustic characteristics. Acoustic analysis is based on the following principle: the acoustic signal is generated in the glottis and passes through the vocal tract where it is modulated by the pharyngeal, buccal, labial and nasal cavities acting as filters [4]. The different frequency bandwidths emitted at the end of the vocal tract are called "formants". The human voice is composed of one fundamental frequency (F0) and four formants (F1 to F4) that correspond to each of the four cavities. Formant frequencies vary over time in function of the spatial conformation changes of the cavities driven by the phonatory muscles. Other characteristics, such as intensity variations and amount of noise contained in the acoustic signal, also are taken into account in the phonetic analysis. Software tools for fast and automated phonetic analysis have been recently developed [5].

The aim of this study was to build and evaluate machine learning models that can recognize OHCA based on the phonetic characteristics of the caller's voice.

2. Methods

The study protocol was approved by the Medical Ethics Committee of Rennes academic hospital (approval number 19.116, issued on December 4, 2019).

2.1. Software

Acoustic features of the recorded calls were extracted with WC-MDX Workstation version 11.6.0.0, UHERS Corporation, 2005. Phonetic analyses were performed with PRAAT v6.1.03, 2019, Institute of Phonetic Sciences, Amsterdam University. All statistical analyses and model building were performed with "R-studio", version 1.3.1093, RStudio PBC, 2009-2021. The following R packages were used: "Dplyr", version 1.0.0, for data manipulation; "MICE", version 3.9.0, for missing data implementation; and "Caret", version 6.0-90, for model building.

2.2. Setting and study population

Data were collected retrospectively from patients for whom a call was done to the emergency call center of Rennes academic hospital, France, between 01/01/2017 and 01/01/2019.

2.3. Variables and groups

The predicted variable was OHCA presence. Patients included in the study were divided in two groups: i) OHCA group if they had OHCA, and ii) NO-OHCA group, if they did not have a diagnosis of OHCA.

Predicting variables were collected by automated phonetic analysis. They were based on the following parameters: fundamental frequency, formants, intensity, jitter, shimmer,

harmonic to noise ratio, number of voice breaks, and number of periods. A full description of all these variables is available in the PRAAT software documentation [5]. Briefly, the spectrogram of the human voice includes five high intensity bands. One corresponds to the fundamental frequency and the other four to formants (figure 1). The *fundamental frequency* (F0) is determined by the tension of the vocal cords and the subglottic pressure, and varies with the stress level. *Formant frequencies* (F1 to F4) are determined by the resonance system volume. *Jitter* measures the short-term variations in the fundamental frequency in seconds, and *shimmer* reflects short-term disturbances in signal intensity. The *noise to harmonic ratio* is defined by the ratio of the non-harmonic and harmonic intensities contained in the acoustic signal. The *number of voice breaks* is the number of distances between consecutive pulses that are longer than a defined duration divided by the pitch floor. The *number of periods* is calculated by counting the number of time intervals between glottal pulses.

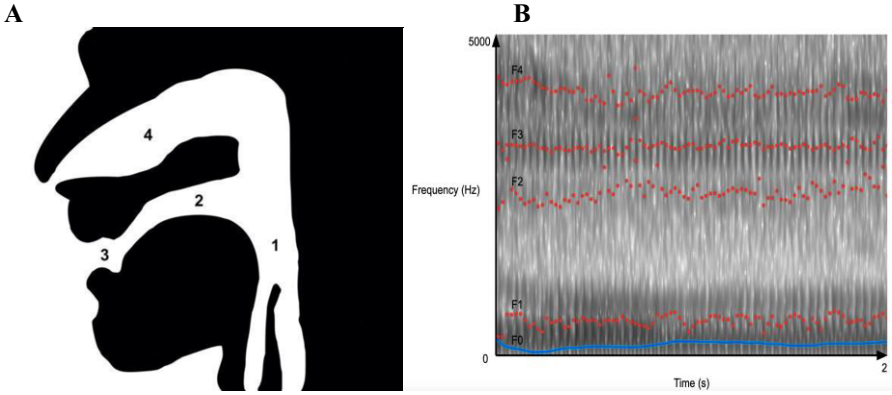


Figure 1: **Bases of the automated phonetic analysis. A: Voice signal.** The acoustic signal is generated in the glottis and modulated by four resonance systems: the pharyngeal (1), buccal (2), labial (3), and nasal (4) cavities. Each cavity acts as a filter, and the bandwidths emitted at the end of the vocal tract are called "formants". **B: Voice spectrogram.** The human voice is composed of one fundamental frequency (F0, blue line) and four formants (F1 to F4, red dots). Formants vary over time according to the spatial conformation changes of the four cavities driven by the phonatory muscles. A spectrogram is the graphic representation of these five bandwidths in function of time.

2.4. Statistics

The Student's *t*-test was used to compare means between groups. A p-value <0.05 was considered significant. To avoid multicollinearity, correlation coefficients were calculated for each pair of variables, and when the coefficient was higher than 0.8, one variable was excluded. *Data splitting*: Data were randomly divided in two parts: training dataset and test dataset. The training dataset corresponded to 80% of the complete dataset and was used to build the models. *Model training*: Three models were constructed: binary logistic regression, random forest (500 trees), and neural network (3 layers). *Performance measurement*: The predictions made by the three models were compared to the "OHCA" variable in the test dataset and receiver operating characteristic curves (ROC) were constructed accordingly. The area under the curve (AUC) was the primary outcome used to evaluate each model performance.

3. Results

3.1. Selected patients

820 patients were included in the study, 410 in each group.

3.2. Predicting variables

Table 1 shows the comparison of the mean values of each selected predicting variable between groups.

Table 1. Comparison of the selected predicting variables between groups. Values in brackets represent the 95% confidence intervals. *t-test significance was set at p <0.05 level. OHCA= Out of Hospital Cardiac Arrest, med=median, min=minimum, max=maximum, sd=standard deviation, n=number, NHR= noise to harmonic ratio.

Variable	OHCA group (n=410)	NO-OHCA group (n=410)	p (t-test)
Pitch mean (Hz)	244 (238 - 250)	197 (192 - 202)	< 0.001*
Pitch sd (Hz)	45.3 (43.5 - 47)	47 (44.9 - 49.1)	0.216
Pitch min (Hz)	136 (131 - 140)	109 (106 - 113)	< 0.001*
Pitch max (Hz)	417 (409 - 425)	395 (386 - 404)	< 0.001*
N of voice breaks	12.2 (11.9 - 12.5)	12.5 (12.2 - 12.9)	0.181
Jitter local absolute	$8.5e^{-5}$ ($8.2e^{-5}$ - $8.9e^{-5}$)	$10.7e^{-5}$ ($10.2e^{-5}$ - $11.2e^{-5}$)	< 0.001*
Jitter RAP (%)	0.82 (0.80 - 0.84)	0.84 (0.81 - 0.87)	0.399
Shimmer local (dB)	1.01 (1.00 - 1.03)	1.03 (1.01 - 1.04)	0.335
Shimmer APQ11 (%)	10.0 (9.7 - 10.2)	9.9 (9.6 - 10.3)	0.930
Mean NHR	0.153 (0.147 - 0.158)	0.154 (0.147 - 0.161)	0.741
Formant med H1 (Hz)	518 (512 - 525)	493 (487 - 498)	< 0.001*
Formant med H2 (Hz)	1482 (1469 - 1495)	1454 (1441 - 1468)	< 0.001*
Formant med H3 (Hz)	2288 (2279 - 2298)	2281 (2271 - 2290)	0.247
Formant med H4 (Hz)	3063 (3057 - 3069)	3074 (3068 - 3081)	0.014*
Intensity med (dB)	68.3 (67.6 – 69.0)	65.2 (64.5 - 65.9)	< 0.001*
Intensity sd (dB)	14.0 (12.9 - 15.2)	11.1 (10.8 - 11.4)	< 0.001*
Intensity min (dB)	27.6 (26.6 - 28.6)	27.7 (27.0 - 28.4)	0.887
Intensity max (dB)	79.6 (79.1 - 80.1)	78.2 (77.7 - 78.8)	< 0.001*

3.3. Model performance analysis

Table 2 shows the performance of the three models.

Table 2. ROC-AUC value for OHCA prediction by each model. Models used the acoustic features of the caller's voice to predict OHCA. Values in brackets represent the 95% confidence intervals.

Model	ROC-AUC
Binary logistic regression	71.4 (63.5-79.4)
Random forest	74.9 (67.4-82.4)
Neural network	64.5 (57.3-71.8)

4. Discussion

This study describes machine learning models that use the acoustic features of the bystander's voice to predict OHCA. These acoustic features were previously identified as stress markers. In 2010, Frampton and al. observed significant variations in pitch and

fundamental frequency in employees facing time restrictions to complete an order [6]. Similarly, Mendoza et al. [7] reported an increase in fundamental frequency, jitter and shimmer perturbations in individuals subject to work-related stressful conditions. To our knowledge, our study is the first to show similar results in the field of emergency healthcare. We observed statistically significant differences in the bystanders' voice acoustic parameters in the presence of OHCA (table 1). We also demonstrated that it is possible to create decision-making support models to recognize OHCA based on these parameters (table 2). These results provides perspectives for short-term applications of machine learning models that integrate semantic and acoustic parameters. Indeed, in 2019, Blomberg et al. developed a model based only on semantic elements that could reduce OHCA detection time [8]. The integration of the acoustic parameters described in the present study should increase the performance of such models.

5. Conclusion

This study demonstrates that machine learning models can recognize OCA based on the acoustic characteristics of the caller's voice. The integration of the acoustic parameters identified in this study could help to increase the performance of decision-making support systems that already integrate semantic parameters in order to improve OHCA detection over the phone.

References

[1] Berdowski J, Berg RA, Tijssen JGP, Koster RW. Global incidences of out-of-hospital cardiac arrest and survival rates: Systematic review of 67 prospective studies. Resuscitation 81, 1479–1487 (2010).
[2] Rossetti AO, Oddo M, Logroscino G, Kaplan PW. Prognostication after cardiac arrest and hypothermia: A prospective study. Ann. Neurol. 67, 301–307 (2010).
[3] Viereck S, Møller TP, Rothman JP, Folke F, Lippert FK. Recognition of out-of-hospital cardiac arrest during emergency calls — a systematic review of observational studies. Scand. J. Trauma Resusc. Emerg. Med. 25, 9 (2017).
[4] Titze IR. Some Consensus has been Reached on the Labeling of Harmonics, Formants, and Resonances. J. Voice 30, 129 (2016).
[5] Praat: doing Phonetics by Computer. https://www.fon.hum.uva.nl/praat/.
[6] Frampton M, Sripada S, Bion RAH, Peters S. Detection of time-pressure induced stress in speech via acoustic indicators.
[7] Mendoza E. Carballo G. Acoustic analysis of induced vocal stress by means of cognitive workload tasks. J. Voice Off. J. Voice Found. 12, 263–273 (1998).
[8] Blomberg SN, et al. Machine learning as a supportive tool to recognize cardiac arrest in emergency calls. Resuscitation 138, 322–329 (2019).

Challenges of Trustable AI and Added-Value on Health
B. Séroussi et al. (Eds.)

doi:10.3233/SHTI220499

An Adverse Drug Reaction Database for Clinical Use – Potential of and Difficulties with the Summary of Product Characteristics

Birgit EIERMANN[a,b,1] , Daniel RODRIGUES[c], Paul COHEN[c] and Lars L GUSTAFSSON[a,c]
[a] *Department of Laboratory Medicine, Division of Clinical Pharmacology, Karolinska Institutet, Karolinska University Hospital, Stockholm, Sweden*
[b] *Inera AB, Swedish Association of Local Authorities and Regions, Stockholm, Sweden*
[c] *Swedish Institute for Drug Informatics (SIDI), c/o Division of Clinical Pharmacology, Karolinska Institutet, Karolinska University Hospital, Stockholm, Sweden*

Abstract. Adverse drug reactions (ADRs) for all drugs in Europe are described in the legally approved Summary of Product Characteristics (SmPC). An overview of all ADRs of the patients' drug list can support healthcare staff to link patient symptoms to possible ADRs. We review the possibilities and challenges to extract ADR information from SmPCs and present the development of our semi-automated procedure for extraction of ADRs from the tabulated section of the SmPCs to create a database, named Bikt, which is regularly updated and used at point of care in Sweden. The existence of five major table formats for ADRs used in the SmPCs required the development of different parsing scripts. Manual checks for correctness for all content has to be performed. The quality of extraction was investigated for all SmPCs by measuring precision, recall and F1 scores (i.e. the weighted harmonic mean of precision and recall) and compared with other methods published. We conclude that it is possible to semi-automatically extract ADR information from SmPCs. However, clear technical and content guidelines and standards for ADR tables and terms from drug registration authorities would lead to improved extraction and usability of ADR information at point of care.

Keywords. Adverse drug reaction; ADR-database, clinical decision support; data extraction; knowledge bases; Summary of Product Characteristics

1. Introduction

Adverse drug reactions (ADRs) are defined by the World Health Organization (WHO) as 'a response to a medicine which is noxious and unintended, and which occurs at doses normally used in man'. ADRs are costly and they are especially common when treating patients with multiple diseases prescribed a high number of drugs. About 80% are dose-dependent and can be avoided [1,2].

[1] Corresponding Author, Birgit Eiermann, Box 17703, 118 93 Stockholm, Sweden; E-mail: birgit.eiermann@inera.se.

In Europe, ADRs for approved drugs are described in section 4.8. of the Summary of Product Characteristics (SmPC) which are approved by the national medical product agencies (MPA) or European Medicines Agency (EMA). EMA publishes guidelines for the SmPC structure for section 4.8 which should contain a tabulated section of ADRs, listed according to their frequency and using the system for organ classification (SOC), where they occur, as defined by MedDRA (Medical Dictionary for Regulatory Activities). However, provided templates for the SmPC offer little guidance for the layout of the table, no support for selection of a controlled vocabulary for ADRs and provide no other guidelines for the description of ADRs. This causes a wide variation in table formats, inconsistent use of terms and inconsistent linking of ADR terms to specific SOC terms with challenges both for machine and human reading [3].

Integrated access to information on ADRs at point-of-care, could assist healthcare staff to choose and dose drugs appropriately for an individual patient. A summary of all the potential ADRs of all patients´ drugs could help physicians by providing a good overview of ADRs to be considered for a patient. An ADR database integrated into an electronic health record system (EHR) could serve as a clinical decision support for healthcare personnel.

This paper outlines the development of processes and methods to build the ADR database "Bikt" in Sweden. It describes various structures of existing SmPC ADR tables, the extraction and updating process of ADR information as well as the procedure for distribution of the database to EHR-systems to be used at point of care. It also points out which table structure is most suitable for automated extraction of ADRs from SmPC documents available in Europe after having performed an experiment. [4]

2. Development of the ADR database and its distribution to EHR systems

The Bikt database contains ADR information for over 95% of all approved drugs on the Swedish market (a total of 5429 SmPCs). It is updated regularly on a monthly basis, distributed centrally, and integrated into several major EHR.

The development process of the ADR database Bikt comprises of the following steps:

- Data import: retrieval of SmPC documents from various websites and their linkage to national drug entities.
- Extraction process: Step 1: transformation of SmPC documents into pure text format with parts of the layout information retained and split into separate text segments based on the chapter and section structure of the SmPC. Step 2: extraction of ADRs using 5 principle parsing scripts. Each SmPC is parsed with each script and the script that gives the best results, is used to classify the ADR table format. The five parsing scripts were developed, in an iterative process using a heuristic method, based on visual inspection of various SmPCs
- Manual editing and approval: a custom-built editing tool is used by pharmacists to control and correct the extracted ADRs following a standardized operation procedure. Additionally, the linkage to the national drug identifier is checked

before approval. Corrections of the extracted material can be e.g removal of terms (e.g. “has been reported”), or footnotes which are not included in Bikt.

- Export and distribution to point of care: Exported ADR tables are integrated into a national database (Sil-database; Sil = Swedish information services for drugs) which is distributed to all counties in Sweden and implemented in all major EHR.
- Update: New SmPCs are added to the database and existing SmPCs are checked for changes in the ADR section.

3. Quality of Data Extraction

In an experiment, we quantified the suitability of the various table formats for correct extraction using the subset of SmPCs which already had been manually checked and approved. We technically compared the corrected subset of SmPCs with a newly extracted version of the same SmPCs. The subset consisted of more than 3200 SmPCs. The performance of our models for extraction of the different table types were evaluated using standard metrics methods: Precision (i.e. true positive/(true positive + false positive)), Recall (i.e. true positive/(true positive + false negative)) and F1-score (i.e. the weighted harmonic mean of precision and recall) [5]. True positive (TP) terms are defined as terms which do exist in the extracted version and in the manually corrected version. False positive (FP) terms are defined as terms which exist in the extracted version but not in the manually corrected version. False negative (FN) terms are defined as terms which do not exist in the extracted version but exist in the manually corrected version. The same applies for the analysis of the combination of SOC and frequency. Two analyses of precision, sensitivity and F1 score were performed to determine best extraction results for the various table formats:

1. Precision, recall and F1 score for “All ADR terms”; which ADR terms are included or missing in the newly extracted version compared to the manually corrected one.
2. Precision, recall and F1 score for “The combination of SOC and frequency”; each SOC is combined with one to many frequencies in an SmPC; this combination is compared in the newly extracted version with the manually corrected one.

During the development of the parsing scripts using an iterative process, we identified 5 table types:

- DFSU - frequencies in the rows followed by indented SOCs and ADRs
- LFSU - frequencies in the column header, and system organ classes followed by ADRs in the rows
- LSFU - SOCs in the column header followed by frequencies and ADRs in the rows
- MSFU - complete tables with SOCs in the rows and frequencies in the column header containing empty table fields where no ADRs exist for certain SOC-frequency combinations
- TSFU - SOCs are displayed per row followed by frequencies and ADRs

(D= Disposition, F = frequency, L = Layout, M = Matrix, S = SOC, T = Table, U = Undesirable effect).
LSFU tables are the most frequent ones used within the SmPC documents (n = 1881).

The results for the extraction of the various table formats are summarized in table 1. The TSFU format (fig. 1) was extracted with the highest precision and recall for both correct ADR terms (0,96 and 0,92, respectively) and the combination of SOC and frequency terms (0,97 and 0,93, respectively).

Table 1. Precision, recall and F1-score calculated for the extraction of adverse drug reaction terms and the combination of system organ class and frequency for all table type variants.

	Extraction of ADR terms			Extraction of combination of SOC and frequency		
Table type (number of SmPCs)	Precision	Recall	F1-score	Precision	Recall	F1-score
DFSU (188)	0,71	0,70	0,70	0,72	0,69	0,7
LFSU (440)	0,85	0,77	0,78	0,41	0,27	0,31
LSFU (1411)	0,66	0,94	0,73	0,92	0,89	0,89
MSFU (549)	0,87	0,58	0,66	0,58	0,43	0,48
TSFU (618)	0,96	0,92	0,93	0,97	0,93	0,94

System organ class	Frequency	Adverse reaction
Immune system disorders	Not known	Hypersensitivity reactions, both local and generalised, including rash*
Psychiatric disorders	Uncommon	Mood swings
Nervous system disorders	Common Uncommon	Headache Dizziness
Vascular disorders	Uncommon	Hot flush

Figure 1. Example of the "table-based" (T) format TFSU: system organ classes (S) are listed in the rows followed by frequencies (F) and undesirable effects (U).

4. Discussion and Conclusions

SmPC documents are important sources of medical information in Europe on approved drugs for healthcare staff. However, these legally approved sources lack standardization of the technical structure as well as of the graphical layout and ADR terminology used. The expressed need of Swedish physicians to have access to an ADR overview resulted in the semi-automated construction of the ADR database, Bikt based on SmPCs of all drugs on the Swedish market and implemented at point of care [6]. Review of the literature showed that there is no other ADR database in Europe being updated regularly and integrated into EHR systems to be used at point of care.

Various table formats led to the development of 5 major parsing scripts. An experiment showed which table format is most suitable for automatic extraction. However, manual

control has to be performed for all SmPCs which is time consuming, costly, prone to introduce mistakes and leads to reduced update frequency. Therefore, the improved use of standards for both table format and terminology within the SmPC guidelines for the ADR tables, and an increased control of the adherence to the standards by the medical product agencies would lead to improvement for both human and machine reading purposes.
Bikt is implemented in over 50% of the counties in Sweden through integration into three major EHR. Future evaluations are planned regarding the use and possible benefits within different groups of healthcare personnel.

Ultimately, information in the SmPCs should be entered according to a defined standard and in a formal and consistently structured way from the beginning to increase their usability as suggested earlier [7]. The development of parsing scripts or other extraction techniques and time-consuming manual control could then be skipped, and information could be used immediately at point of care.

References

[1] Lagnaoui R, Moore N, Fach J, Longy-Boursier M, Bégaud B. Adverse drug reactions in a department of systemic diseases-oriented internal medicine: prevalence, incidence, direct costs and avoidability. Eur J Clin Pharmacol 2000;56:181-6.
[2] Natanaelsson J, Hakkarainen KM, Hägg S, Andersson Sundell K, Petzold M, Rehnberg C, Jönsson AK, Gyllensten H. Direct and indirect costs for adverse events identified in medical records across care levels, and their distribution among payers. Res Social Adm Pharm 2016; pii: S1551-7411:30237-6.
[3] Lamy JB, Ugon A, Berthelot H. Automatic extraction of drug adverse effects from product characteristics (SPC): a text versus table comparison. Stud Health Technol Inform 2016;228: 339-43.
[4] Eiermann B, Rodriguez D, Cohen P, Gustafsson LL. ADR databases for on-site clinical use: potentials of Summary of Products Characteristics. Basic Clin Pharmacol Toxicol. 2021; 28(4):557-567.
[5] Van Rijsbergen C. Information retrieval. Second edition London, UK: Butterworths; 1979.
[6] Rahmner PB, Eiermann B, Korkmaz S, Gustafsson LL, Gruvén M, Maxwell S, Eichler HG, Vég A. Physicians' reported needs of drug information at point of care in Sweden. Br J Clin Pharmacol 2012;73:115-25.
[7] Maxwell S, Eichler HG, Bucsics A, Haefeli WE, Gustafsson LL; e-SPC Consortium. e-SPC - delivering drug information in the 21st century: developing new approaches to deliver drug information to prescribers. Br J Clin Pharmacol 2012;73:12-5.

Challenges of Trustable AI and Added-Value on Health
B. Séroussi et al. (Eds.)

doi:10.3233/SHTI220500

Support-Vector Machine-Based Classifier of Cross-Correlated Phoneme Segments for Speech Sound Disorder Screening

Emilian-Erman MAHMUT[a,1], Stelian NICOLA[a] and Vasile STOICU-TIVADAR[a]
[a] *Politehnica University Timisoara, Romania*
Department of Automation and Applied Informatics

Abstract. This paper presents a Support-Vector Machine (SVM) based method of classification of cross-correlated phoneme segments as part of the development of an automated Speech Sound Disorder (SSD) Screening tool. The pre-processing stage of the algorithm uses cross-correlation to segment the target phoneme and extracts data from the new homogeneously trimmed audio samples. Such data is then fed into the SVM-based classification script which currently achieves an accuracy of 97.5% on a dataset of 132 rows. Given the global context of an increasing trend in the incidence of Speech Sound Disorders (SSDs) amongst early-school aged children (5-6 years old), the constraints imposed by the new Corona virus pandemic, and the (consequent) shortage of professionally trained specialists, an automated screening tool would be of much assistance to Speech-Language Pathologists (SLPs).

Keywords. Speech Sound Disorders, Support-Vector Machine, cross-correlation

1. Introduction

Phonemes are the smallest speech sound units capable of changing meaning in a language. They play a central role in the structure of speech and have the potential to transcend all linguistic tiers of human languages, from phonetics all the way through to semantics and beyond (pragmatics/phraseology). Phonemes are a major acquisition in the early stages of both generating and decoding human speech. As substantiated in papers [1] and [2], it is highly likely that such stages rely on statistical models.

Phonological assimilation is a phenomenon whereby in any given word, the pronunciation of a phoneme is affected by the neighboring phonemes. It produces its effects both progressively, when the phoneme *n* is affected by the phoneme *n+1*, and regressively, i.e., the phoneme *n+1* is affected by the phoneme *n*. Mathematical modelling of speech sounds can be particularly challenging given that regressive phonological assimilation seems to operate backwards in time. Paper [2] hints at infants (and adults) being able to compute "both forward and backward conditional probabilities to segment continuous artificial streams" and the neuroscience perspective [3] confirms

[1] Emilian-Erman Mahmut, Politehnica University Timisoara, P-ta Victoriei no. 2, Timisoara, Romania; E-mail: emilian.mahmut@aut.upt.ro

such considerations and acknowledges the shortcomings of the conventional fixed frame size and rate segmentation technique commonly used in automatic speech recognition (ASR).

Paper [4] provides a presentation of the general framework of automated speech segmentation and a thorough review of various segmentation algorithms and feature extraction techniques.

According to the criteria initially adopted for its development, the SSD screening tool should: be language independent (mathematical model), provide real-time or near-real-time feedback (low computational cost), use open-source software (low financial cost), be web-based, and use mobile technology, grant easy and open access to SLPs/researchers and subjects' parents. Similar research projects aimed at providing an automated phoneme classification [5,6,7], reviewed in paper [8] reported classification accuracy rates roughly comprised between 78 and 86%.

This paper presents a method for the automated SSD screening in early-school aged subjects, which generates better results than the known literature.

2. Method

The diagram below provides a synthesis of the 3-stage automated processing of the .wav files: the pre-processing stage consisting of the cross-correlation (X-corr) bases segmentation followed by the feature extraction stage, and the SVM-based classification stage (Figure 1). Python scripts were written to perform the processing stages described above.

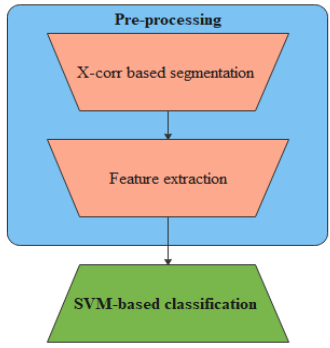


Figure 1. Processing diagram

2.1. Pre-processing

2.1.1. X-corr based segmentation

A total number of 132 .wav files were processed: subjects' pronunciations of words containing the target phoneme /r/ in medial position. The subjects were 5–6-year-old pupils whose pronunciations were recorded during the traditional screening process performed by the SLP of the CNB school in Timisoara (Romania) in 2019.

While segmenting (and extracting features from) a word-initial or a word-final phoneme provides a natural boundary – because of the silence before and after the word

– a rigorous segmentation of a medial target phoneme entails a more complex approach. The words fed into the algorithm were *mere*, *pere*, and *mură* (Romanian words for *apples*, *pears*, and *blackberry*), containing the /r/ phoneme in between vowels. The algorithm was devised to generate homogeneous medial-phoneme segments by cross correlating each of the subjects' utterances with the SLP (reference) pronunciation and it is presented in detail in paper [9]. Providing optimal alignment of the 2 audio signals, cross-correlation solves the issue of the variable length of a pronunciation of the same word by different people. It already starts on the classification problem since it generates homogeneously trimmed target-phoneme segments, and it ensures a smooth transition into the classification stage. The features extracted from the target-phoneme segments, briefly described in the following subsection, are used as input for the classification stage.

2.1.2. Feature extraction

The pre-processing stage results in a 10-column table containing: *Max_left, Max_right, Left_index, right_index, Maxamp_SLP, Maxamp_SUB, Infl_SLP, Infl _SUB, R-squared, Coefficient of determination* (Table 1a and 1b). An 11th column was added to the table to indicate the SLP Opinion expressed as boolean label. *Max_left/Max_right* is the maximum cross-correlation coefficient found between the 2 audio signals by shifting the sample signal (SLP) to the left and, respectively, to the right of the reference signal (Subject). *Left_index/Right_index* are the indices where the maximum cross-correlation coefficient was found. *Maxamp_SLP* and *Maxamp_SUB* report the maximum amplitude value detected in the SLP and the Subject pronunciation of the cross-correlated phoneme segments. *Infl_SLP/Infl_SUB* indicate the total number of inflections (Gaussian smoothing, 5% standard deviation) in each of the cross-correlated target-phoneme segments.

2.2. SVM-based Classification

The features of interest are extracted into a .csv file and are divided into 2 categories: training data and test data. A choice was made to assign 70% of such data to the training data category while the remaining 30% are designated as test data. The main libraries used in the Python script are *pandas*, *numpy*, *sklearn*, and *matplotlib*.

Table 1a. Feature extraction table (first 5 columns)

Subject No.	Max_left	Max_right	Left_index	Right_index	Maxamp_SLP
1	15.57	**31.11**	-1865	998	4503.00
2	17.91	**18.41**	-418	141	7328.00
3	**20.60**	19.22	-99	93	8128.00
4	**52.30**	37.35	-1493	172	7982.00
…					
131	48.35	**74.14**	-503	1755	5469.00
132	**57.74**	42.73	-4836.00	5077.00	6839.00

Table 1b. Feature extraction table (last 6 columns)

Maxamp_SUB	Infl_SLP	Infl_SUB	R^2	Coefficient of determination	SLP Opinion
3337.00	219.00	160.00	0.99	0.07	0
3682.00	168.00	168.00	0.98	0.96	0
10089.00	189.00	185.00	0.99	0.99	1
7765.00	180.00	188.00	0.99	0.99	1

…					
8044.00	174.00	200.00	0.99	0.90	0
12518.00	172.00	182.00	0.99	0.98	1

The Python SVM-based classification algorithm processes the 2 data categories and reports the following metrics: accuracy, precision, recall, and the F1 score.

3. Results

In 68 target-phoneme segments, the segmentation algorithm detected the maximum cross-correlation coefficient (*Max_left*) by shifting the sample signal (Subject's pronunciation) to the left of the reference sample (SLP's pronunciation) while the remaining 64 phoneme segments were generated by shifting the sample signal to the right of the reference sample. The phonetical implication is that 68 subjects needed more time to transition from the preceding vowel and achieve the full vibration of the liquid consonant /r/ in correlation with the reference signal. The other 64 subjects needed less time, as compared to the reference signal, to transition to the target phoneme. Table 2 below illustrates the classification metrics for the 132 .wav files analyzed in this paper.

Table 2. SVM metrics

Accuracy	Precision	Recall	F1 score
0.975	1.00	0.928	0.962

To better illustrate the results, a Confusion Matrix was generated (Figure 2).

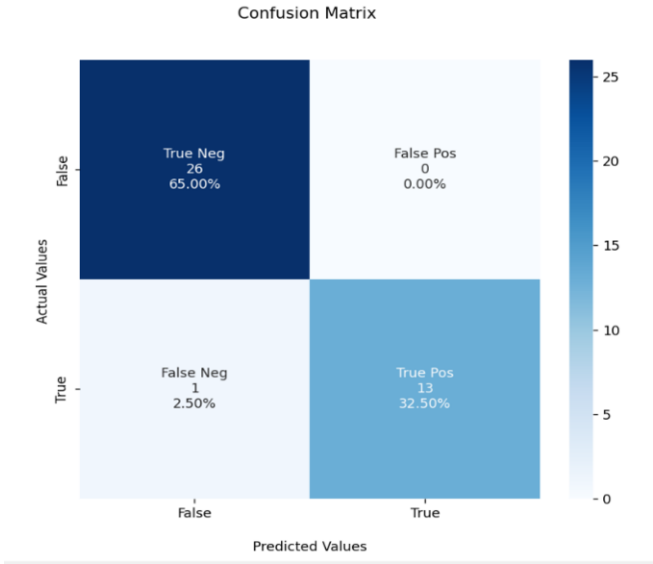


Figure 2. Confusion Matrix

As synthesized in the confusion matrix, 65% of the test data were classified as true negatives (TN), 32.5% are true positives (TP), 2.5% are false negatives (FN) and there were no false positives (FP).

4. Discussion and conclusions

Early diagnosis of SSDs (before the age of 7) is crucial since it allows for speech therapy sessions to be administered with optimal results, preventing therefore the onset of stigma that may lead to serious behavioral issues in early school-aged subjects [10]. An unbiased verdict and the anonymity provided by a web-based/mobile-technology solution would help to tackle uninformed parents' adversity to speech therapy/pathology. The current results are promising, the accuracy achieved is better than the rates reported in the literature.

The *keras* library (TensorFlow framework) was used to configure an artificial neural network in order to compare and validate the metrics described above. Such neural network requires a lot more data rows (in the order of thousands) than the ones currently available so as to generate reliable results. Paper [11] contains a detailed discussion on classification algorithms and proposes a voting classifier to aggregate the predictions of a range of individual classifiers. Other phonemes in various vowel/consonant contexts are being segmented and will be fed both into the SVM-based and into the neural network-based classification scripts.

The final objective of the research project is the development of an application meant to assist the SLPs in their laborious and time-consuming activity, to build a database for interdisciplinary research, and to grant parents access to an unbiased SSD screening tool.

References

[1] Gervain J. Neurocognitive Development: Normative development. In: Handbook of clinical neurology. 2020.

[2] Federmeier K, Huang H-W. Adult and second language learning. In: The psychology of learning and motivation; 2020 May. vol. 72.

[3] Lee B, Cho KH, Brain-inspired speech segmentation for automatic speech recognition using the speech envelope as a temporal reference. Scientific Reports 6, article number: 37647; 2016 Nov.

[4] Alaa ES, Sherif MA, Salah EH, Mohsen R. A Review: Automatic speech segmentation, International Journal of Computer Science and Mobile Computing, 2017 Apr; 6(4):308-15

[5] Grigore O, Grigore C, Velican V. Intelligent system for impaired speech evaluation. Recent Advances in Circuits, Systems and Signals. Book Series: International Conference on Circuits Systems Signals, 2010; Malta; p. 365-8.

[6] Grigore O, Velican V. Self-organizing maps for identifying impaired speech. In: AECE (Advances in Electrical and Computer Engineering); Romania; 2011 Jan; 11(3): 41-8.

[7] Yin S-C, Rose R, Saz O, Lleida E. A study of pronunciation verification in a speech therapy application. In: IEEE International Conference on Acoustics, Speech and Signal Processing; 2009 Apr 19-24; Taipei, Taiwan; p. 4609-12.

[8] Mahmut EE, Stoicu-Tivadar V. Current challenges in the computer-based assessment of speech sound disorders. In: IEEE 12th International Symposium on Applied Computational Intelligence and Informatics (SACI); 2018 May 17-19; Timisoara, Romania; p. 431-6.

[9] Mahmut EE, Nicola S, Stoicu-Tivadar V. Cross-correlation based automatic segmentation of medial phonemes. In: International Symposium on Electronics and Telecommunications (ISETC); 2020 Nov 5-6; Timisoara, Romania; p. 1-4.

[10] Shahin M, Ahmed B, Smith DV, Duenser A, Epps J. Automatic screening of children with speech sound disorders using paralinguistic features. In: IEEE 29th International Workshop on Machine Learning for Signal Processing (MLSP); 2019. p. 1-5.

[11] El-Kenawy E-S M, Ibrahim A, Mirjalili S, Eid MM, Hussein SE, Novel feature selection and voting classifier algorithms for COVID-19 classification in CT images. In: IEEE Access; 2020. vol. 8, p. 179317-35.

Challenges of Trustable AI and Added-Value on Health
B. Séroussi et al. (Eds.)

doi:10.3233/SHTI220501

Speak-PIM, Towards a Framework for the Automatic Detection of Potentially Inappropriate Prescriptions

Abdelmalek MOUAZER[a,1], Karima SEDKI[a], Rosy TSOPRA[b,c,d] and Jean-Baptiste LAMY[a]
[a] *Université Sorbonne Paris Nord, LIMICS, INSERM, F-93000, Bobigny, France*
[b] *INSERM, Université de Paris, Sorbonne Université, Centre de Recherche des Cordeliers, F-75006 Paris, France*
[c] *HeKA, INRIA Paris, France*
[d] *Department of Medical Informatics, Hôpital Européen Georges-Pompidou, AP-HP, Paris, France*

Abstract. Potentially inappropriate medications (PIMs) have adverse health consequences, particularly in elderly patients. Various explicit criteria have been developed to detect PIMs. However, it is difficult to apply these criteria without the help of an electronic decision support tool. Programming these tools can be very complex. Indeed, for computer scientists it is difficult to understand medical issues and for clinicians it is difficult to program in a computer programming language. In this work we present Speak-PIM, a framework for formalizing the PIM's rules. Speak-PIM is based on a very simple semantics which is suitable for the declaration of PIMs without embarking on all the complexity of description logic or computer languages. It aims to offer an efficient collaboration between the computer scientists and clinicians.

Keywords. Therapeutic guidelines, Potentially inappropriate medications, Terminology, Computer programming, Decision Support System.

1. Introduction

Appropriate prescribing for the elderly is a challenge for prescribers. Several assessment tools are available. The explicit tools provide recommendations to avoid over-prescription of drugs that are not clinically indicated, omission of drugs that are necessary and incorrect prescriptions of drugs that may be indicated. The term "Potentially Inappropriate Medicines (PIMs) for the Elderly" is used to refer to medicines that should not be prescribed to this population.

Existing tools for detecting PIMs are clinical practice guidelines in natural language and include several rules. In practice, it is time-consuming for the clinician to review all these rules for each drug and for each patient. Thus, there is a great interest in implementing a computer program for automatically executing the rules.

1 Corresponding Author, Abdelmalek MOUAZER, Université Sorbonne Paris Nord, LIMICS, INSERM, F-93000, Bobigny, France; E-mail: malikmouazer@gmail.com

The passage of the text to a formalized language understandable by the computer poses several problems: (i) the understanding of the text which requires medical knowledge to understand the meaning and overcome the ambiguities, (ii) the difficulties of formalizing the medical language which requires computer and medical computer skills, (iii) the implementation of the rules in a computer language which requires computer skills, (iv) the updating of the rules can be frequent regarding the evolution of knowledge. Many different types of experts are needed to solve all these problems which can generate considerable costs. We hypothesize that a rule-building system could overcome these problems. Some approaches have been proposed in the literature [1][2]. They are presented as formalization systems for all types of clinical practice guidelines, which means that they retain great complexity.

In this work, we propose Speak-PIM, a framework to formalize the rules found in PIMs guidelines, with a simplified syntax easily understandable for clinicians whether for formalization or for validation, and a formal semantics allowing its translation into SPARQL queries.

2. Methods

2.1. Analysis of PIM guidelines

The design of the framework is first preceded by the analysis of the logic of the rules detecting the PIMs (rules found in the PIMs guides). At this stage, we analyzed the STOPP & START guideline [3]. We have taken this guideline as a reference because it is widely used in the literature and because it includes more complex rules, compared to other guidelines, some of them being limited to a list of unconditionally inappropriate drugs. The purpose of this step is to identify the logic elements needed for detecting PIMs (such as logical operations), the clinical elements and the attributes necessary for their declaration.

The logic of expression of the criteria:
In all STOPP & START guidelines, each rule is formulated as follows:

In the case of a STOPP rule: a prescription is considered potentially inappropriate if it is present, possibly with some additional conditions. The conditions may be the presence and/or absence of one or more clinical elements. These elements can be mandatory (succession of logical “and” between these elements), or one of these elements is enough to declare a prescription as being potentially inappropriate (combination of elements with logical “or”). For example, stopp_H1 rule states : “*Stop non-steroidal anti-inflammatory drug (NSAID) other than COX-2 selective agents with history of peptic ulcer disease or gastrointestinal bleeding, unless with concurrent PPI or H2 antagonis*t”. To declare NSAID as being a PIM, the patient must present at least *gastrointestinal bleeding* or a have an history of peptic ulcer (presence peptic ulcer history "or" *gastrointestinal bleeding*) and at the same time his overall treatment does not include PPI or h2 antagonists (absence of PPI “and” absence of H2 *antagonis*t).

In the case of a START rule: the recommended prescription must be absent from the current drug order, and some additional conditions must be satisfied. The conditions are expressed in the same way as for STOP rules.

2.2 .Coding and mapping

Once the rule is declared with Speak-PIM, it is automatically translated into a SPARQL query using a program written in Python.

The input data model is based on the OMOP-CDM [4] model translated in a previous work in OWL.

Several terminologies can be used to code clinical elements when using Speak-PIM. For simplicity, the terminologies used in this work are: ATC for prescriptions, ICD10 for diseases and LOINC for observations. These terminologies have been managed by Owlready2 [5].

3. Results

3.1 Presentation of the framework

The formalization of a rule with the SPEAK-PIM framework must be done in two steps: the declaration of the clinical elements and the writing of the rule.

The declaration of the clinical elements: The first step to formalize a rule with SPEAK-PIM consists in declaring the clinical elements necessary for the expression of the rule with the framework. There are three categories of clinical elements: prescriptions, diseases, and observations (results of biological tests). Each element can be defined by a set of attributes which can be mandatory or optional. Table 1 shows the different attributes available for defining a clinical element.

Table 1. List of possible attributes for each clinical element

Prescription	Disease	Observation
-concepts -indication -is_ongoing -duration -dose_unit -lower_than -greater_than	**-concepts** -is_active	**-concepts** -value -lower_than -greater_than

Concepts: terminological code(s) of the clinical element; indication: terminology code(s) of the disease for which the drug was prescribed; is_ongoing: if the prescription is in progress (true or false); duration: treatment duration in days; dose_unit: dose expression unit; lower_than(Prescription): lower critical dose threshold per day; greater_than(Prescription): upper critical dose threshold per day; is_active: if the disease is current or it is an antecedent; value: observation value; lower_than(Observation): critical lower threshold of the observation; greater_than(Observation): critical upper threshold of the observation.

The Form of the rule: Each rule is in the form:

RULE_XX = Check_pim(action, target, presents = [EP1, EP2, ... EPn One_among(EPO1, EPO2, ... EPOn)], absents =[EA1, EA2, … EAn])

arguments:

action: can take two values “STOP” to stop a prescription or “START” to start a new prescription**.**

target: target prescription, this is the prescription to be stopped or started.
presents = [Epi, One_among(EPOi)]: all the clinical elements in this list must be present to trigger the rules. If only one of a set of clinical elements is sufficient to trigger the rule, all of these elements are put in the "One_among()" object. Between all the objects in the "present" list there are logical connectives "AND".
absents = [EAi]: all the clinical elements in this list must be absent to trigger the rule.

Rule execution: The execution of the rule, namely: carrying an action (stop or start) on a target prescription is carried out when the condition (present AND absent is true). In a more formal way, we can write the following formula:

$$[EP1 \wedge EP2 \wedge ... \wedge EPn \wedge (EPO1 \vee EPO2 \vee ... \vee EPOn)] \wedge [\neg EA1 \wedge \neg EA2 \wedge ... \wedge \neg EAn] \Rightarrow action(target)$$

3.2 Application to STOPP & START CRITERIA

To test the capacity of this framework to formalize PIMs, we used it to formalize the STOPP & START criteria. These criteria have previously been translated into a structured format [6]. We used this work as a basis for comparison and pre-validation of Speak-PIM. Figure 1 shows an example of formalization of the "stopp D4" criterion from the STOPP & START criteria.

```
texte_stopp_d4 = """Stop selective serotonin re-uptake inhibitors (SSRI's)
with current or recent significant hyponatraemia i.e. serum Na+ < 130 mmol/l
(risk of exacerbating or precipitating hyponatraemia)."""hypo_na =
Disease(concepts=["E87.1"])hypo_na_mesure =
Observation(concepts=["77139-4"], lower_than(130))ssris =
Prescription(concepts=["N06AB", "N06CA03"])rule_D4 = Check_pim(
action = "STOP", target = ssris), presents = [ssris, One_among(hypo_na,
hypo_na_mesure)]
)
```

Figure 1. Example of formalization of the stopp D4 criterion from the STOPP & START criteria

Of the 80 STOPP rules, 77 rules were formalized; (A1, A2 and A3 have not been formalized because they are very general and do not target specific drugs). The 77 formalized rules required 83 rules with Speak-PIM. Indeed, a few guideline rules had high-level disjunctions and needed two Speak-PIM rules to be formalized. Moreover, all 34 START rules have been formalized with 41 rules with Speak-PIM.

Applied to a test patient, all the STOPP & START rules could be applied in 0.05 seconds (Dell Inc. Latitude 7410; 15,3 Gio; Intel® Core™ i7-10810U CPU @ 1.10GHz × 12).

4. Discussion and Conclusions

The framework we present is easy to use and understand, whether for the formalization of PIMs guidelines or for their validation. It makes it possible to target PIMs with taking into account the patient's conditions, which is important to avoid "alert-fatigue". The

execution time of all the rules is satisfactory for a real-time implementation in a computer application. Speak-PIM offers a limited and rigid syntax different from standard reference syntax like an "Arden syntax" which offers greater freedom of expression of the rules. This rigidity is desired to allow health professionals to be able to express the rules of PIMs guidelines without computing knowledge.

The data model based on OMOP-CDM makes it interoperable with all patient record systems based on this model without recourse to mapping between different models.

Speak-PIM does not take into account the severity of diseases or the effectiveness of a drug on a patient, information that is sometimes a decisive criterion for declaring a drug as PIM. Technically, we can add an attribute (efficiency, severity of the disease, etc.) but in practice this piece of information is missing from patient records.

Currently, Speak-PIM offers limited support for drug dose. Indeed, our model can only capture the quantity of an active ingredient administered per day, but the characterization of the dosage is more complex than what we propose. However, in the PIM guidelines, all the doses identified were relative to the quantity administered per day.

In the future, we plan to translate Speak-PIM into other computer languages.

Acknowledgments

This work was partly funded by the French Research Agency (ANR) through the ABiMed project [grant number ANR-20-CE19-0017-02].

References

[1] Kluegl P, Toepfer M, Beck PD, Fette G, Puppe F. UIMA Ruta: Rapid development of rule-based information extraction applications. Natural Language Engineering. 2016 Jan;22(1):1-40.

[2] Lamy JB, Ebrahiminia V, Riou C, Seroussi B, Bouaud J, Simon C, Dubois S, Butti A, Simon G, Favre M, Falcoff H. How to translate therapeutic recommendations in clinical practice guidelines into rules for critiquing physician prescriptions? Methods and application to five guidelines. BMC medical informatics and decision making. 2010 Dec;10(1):1-4.

[3] O'Mahony D, O'Sullivan D, Byrne S, O'Connor MN, Ryan C, Gallagher P. STOPP/START criteria for potentially inappropriate prescribing in older people: version 2. Age and ageing. 2014 Oct 16;44(2):213-8.

[4] Lamy JB, Mouazer A, Sedki K, Tsopra R. Translating the Observational Medical Outcomes Partnership-Common Data Model (OMOP-CDM) electronic health records to an OWL ontology. InMEDINFO 2021.

[5] Lamy JB. Owlready: Ontology-oriented programming in Python with automatic classification and high level constructs for biomedical ontologies. Artificial intelligence in medicine. 2017 Jul 1;80:11-28.

[6] Huibers CJ, Sallevelt BT, de Groot DA, Boer MJ, van Campen JP, Davids CJ, Hugtenburg JG, van Hout HP, van Marum RJ, Meulendijk MC. Conversion of STOPP/START version 2 into coded algorithms for software implementation: a multidisciplinary consensus procedure. International Journal of Medical Informatics. 2019 May 1;125:110-7.

Challenges of Trustable AI and Added-Value on Health
B. Séroussi et al. (Eds.)

doi:10.3233/SHTI220502

CDS-Compare: A Web Application for Machine Learning Assisted Curation of Clinical Order Sets

Zackary FALLS[a], Steven H BROWN[b,c], Naveed SHAH[a], Frank LeHOUILLIER[a], Randeep BADWAL[a], Kendria HALL[a], Gillian FRANKLIN[a], William H KELLY[a], Andrew HOLDAWAY[b], Diane MONTELLA[b], Linda WEDEMEYER[b], Sarita KENI[b], Eric ROSE[b,d], Jonathan R. NEBEKER[b,e] and Peter L ELKIN[a,b,f,g,1]

[a] *Department of Biomedical Informatics, Jacobs School of Medicine and Biomedical Sciences, University at Buffalo, USA*
[b] *Department of Veterans Affairs, Office of Health Informatics, USA*
[c] *Department of Biomedical Informatics, Vanderbilt University Medical Center, USA*
[d] *Department of Biomedical Informatics and Medical Education, University of Washington, USA*
[e] *Department of Medicine, University of Utah, USA*
[f] *Department of Veterans Affairs, WNY VA, Bioinformatics Laboratory, USA*
[g] *Faculty of Engineering, University of Southern Denmark, Denmark*

Abstract. Order sets that adhere to disease-specific guidelines have been shown to increase clinician efficiency and patient safety but curating these order sets, particularly for consistency across multiple sites, is difficult and time consuming. We created software called CDS-Compare to alleviate the burden on expert reviewers in rapidly and effectively curating large databases of order sets. We applied our clustering-based software to a database of NLP-processed order sets extracted from VA's Electronic Health Record, then had subject-matter experts review the web application version of our software for clustering validity.

Keywords. Order sets, Database Curation, Clinical Decision Support, Machine Learning

1. Introduction

Clinical order sets, collections of orders for a clinical scenario, are one aspect of clinical decision support that has been shown to improve clinician efficiency and patient safety [1,2]. Each clinical site may have a unique aggregation of orders into an order set for any given condition or situation, and certain sites may use different order sets with more frequency due to population factors, such as prevalence of certain diseases [2]. The use of order sets has also been shown to reduce the variation in patient care for a single site; however, order sets vary across different sites, so variation in care exists from site to site [2]. If the goal is to adopt guideline-based order sets, then they should be standardized across multiple sites [3]. This is especially true for large hospital networks like the US

[1] Corresponding Author, Peter L. Elkin, Buffalo, NY, USA; E-mail: elkinp@buffalo.edu.

Department of Veterans Affairs (VA) and its nationwide network of medical facilities. However, creating order sets is time and resource intensive [2]. If we can reduce the difficulty in order set curation within and between facilities, then we can take steps towards a system-wide database of patient-centered order sets that adhere to condition-specific guidelines, ultimately reducing unwanted variation in patient care across all VA medical facilities [4,5]. We further anticipate that this ability will help smooth VA's transition from 128 legacy CPRS installations to a single Cerner Millennium instance shared with the US Department of Defense by reducing multi-site order sets to a single, well-curated dataset (https://www.ehrm.va.gov/resources/factsheet) [6].

This study describes the clustering performance of a clustering method in the pragmatic setting of a web application, CDS-Compare, that will enable experts to curate clinical order set databases. We have created clinically relevant clusters using a hierarchical clustering method and an effective Natural Language Processing (NLP) algorithm to create features for each order set based upon standardized terminologies. This methodology was applied to a VA order set dataset of medication orders and embedded within a user-friendly web application for end-user experts to access and rapidly curate the dataset. To evaluate our software's performance and usability, we asked ten clinicians to demo our web application, assess the clustering efficacy by comparing inter- and intra-cluster order sets, and then fill out a questionnaire with pointed questions regarding the usability of the interface.

2. Methods

2.1. Order sets extraction and featurization

The database of order sets used in this study were extracted from the VA's Electronic Health Record, a component of the VA's Electronic Health Record Modernization initiative. The dataset used herein consisted of 1,293 order sets with an average number of orders/per order set of 12.6 (minimum = 1, maximum = 141). This dataset was a subset of prescription medication order sets and was not the complete set of order sets across all VA medical sites. The complete dataset will contain both medications and procedures.

Each order set consists of one or numerous medication orders,. We wanted a way to compare all of the order sets by the orders which they contained, therefore we applied High-Definition Natural Language Processing (HD-NLP), a well-tested, high-speed system that processes clinical data and extracts clinically relevant terms in a standardized form, e.g., SOLOR, SNOMED-CT, RxNorm, and LOINC [7,8]. Processing of the study order sets with HD-NLP resulted in 860 unique SNOMED-CT, RxNorm, and LOINC codes across the 1,293 order sets present in our dataset, and every order was mapped to at least one code. We used the 860 unique codes extracted with HD-NLP as features to represent each of the order sets as a bit vector of length 860 with each bit representing the presence or absence of a code.

2.2. Hierarchical clustering

We leveraged an agglomerative hierarchical clustering algorithm to enable easier and more rapid curation of order sets. We determined the optimal cosine distance threshold using the elbow heuristic method which results in the clusters that best fit the data. After computing clusters for the order sets, we wanted to provide the end-users with a useful

visualization of the order sets and their corresponding clusters. We used the dimensionality reduction method t-SNE (t-distributed stochastic neighbor embedding) to reduce the 860-dimension vectors to two dimensions allowing for easier visualization.

2.3. Web application

The last stage of our software development was the creation of the web application interface, where users will take advantage of the clustering method employed to curate the dataset of order sets more easily. The first component of the software, a table containing all order sets, provides the user with the specific name of each order set and the cluster to which it belongs so the user can use both pieces of information to isolate order sets or groups of order sets that may need to be assessed together. Second is the t-SNE plot (Figure 1) , which is embedded into the web interface as a 2D visualization of the order sets and their clusters to allow the user to select specific order sets for view in the comparison tables. All or user-selected order sets are populated in the plot with the marker color-coded based upon the cluster to which the order set belongs. This enables users to easily identify order sets that are most similar by color and relative location to one another.

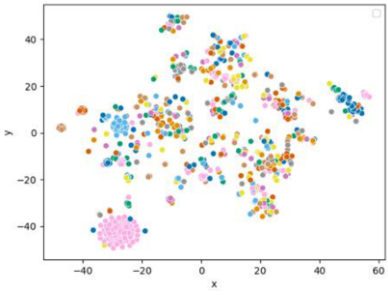


Figure 1. t-SNE plot for visualization of clusters of order sets where each color represents a unique cluster.

Lastly, the side-by-side comparison tables are where users can view two order sets at a time with all the orders and pertinent information corresponding to those order sets. This component is critical for users to determine if there is redundancy in order sets or large overlap where two order sets can be merged to create more robust order sets for clinical use. HD-NLP extracted codes used for clustering are not presented to the user.

Ten clinicians assessed the application clustering performance in CDS-Compare by selecting five order set pairs where both order sets in each pair belonged to the same cluster, and determined the similarity between the two order sets in each pair (intra-cluster comparison) based on the orders (prescription medications). The clinicians assessed each order set pair using three mutually exclusive choices: “same” - order set pairs were 100% identical; “different” - order set pairs were unique with minimal to no overlap in orders; and “overlapping” - order set pairs had substantially similar orders. The same ten clinicians then repeated this procedure on five order set pairs where any two order sets in each selected pair belonged to different clusters (inter-cluster comparison). We state that the clustering works well if order sets are the same or highly overlapping intra-cluster significantly more often than order sets in different clusters.

3. Results

The ten reviewers in all analyzed 50 intra-cluster order set pairs and 50 inter-cluster order set pairs. The cluster validity results showed 66% of the intra-cluster order set pairs assessed had overlap and 22% had the same orders. The inter-cluster assessment showed 72% of order set pairs were different and 28% had overlap. The intra-cluster analysis showed 44/50 were either the same or overlapping, while the inter-cluster analysis showed 14/50 order set pairs were overlapping ($p<0.001$; Pearson Chi-Square test).

We asked the reviewers to rate aspects of their subjective experience with the web application by indicating the degree to which they agreed or disagreed with 10 statements using a Likert scale of 1 to 5 (1: strongly disagree, 2: disagree, 3: neutral, 4: agree, 5: strongly agree); results are aggregated in Table 1. They found that order set contents were easy to view (average score of 4.40), they were able to make judgments about order set content (4.40), and the web application was both easy to access (4.40) and visually pleasing (4.50). Further development of CDS-Compare is needed for clinical understanding (4.00) and consistency across reviewers (3.90) .

Table 1. Questions and corresponding results from our Likert scale questionnaire.

Questions	Average score	Median score	Minimum score	Maximum score
The system was easy to access?	4.40	5	2	5
The system was easy to use?	3.90	4	2	5
You could view the contents of the order sets?	4.40	5	2	5
You could make judgments about the content of the order sets?	4.50	4.5	4	5
The system was visually pleasing?	4.50	5	3	5
The ability to focus on certain topics was helpful?	4.10	4	3	5
The system was easy to understand from a clinical perspective?	4.00	4	2	5
The system had the functionality I needed?	4.10	4	2	5
The system in my opinion could be used consistently across reviewers?	3.90	4	2	5
I enjoyed using the system?	4.30	5	2	5

4. Discussion

Analysis of the cluster validity assessments show that the software is clustering the order sets in a way that effectively groups order sets that are the same or contain highly overlapping orders into the same cluster. Moreover, the order sets in different clusters were significantly less likely to be overlapping. These results support the use of this application to reduce the difficulty for experts to select and remove redundant order sets in the database. Use of this application by experts will result in a highly curated set of order sets that will reduce variation in patient care across VA medical facilities.

With regards to the usability of the web interface, the overall responses were very positive. Many aspects of the web interface were well liked by the reviewers, such as ease of access and ability to make judgments about the order sets. We still need to further develop components of the interface to enhance clinical understanding of the order sets for users and improve potential inter-reviewer consistency. Constructive comments provided by the reviewers also alluded to some of the needed improvements, such as more clinically relevant variables for the orders and clearly written instructions on use of the application. We continue to improve the interface using clinician feedback.

5. Conclusion

Based on the results obtained from our reviewer assessment of the clustering and web interface for CDS-Compare, we have a promising web application for curation and accurate clustering of large order set databases. The clustering method used within the application is effective at grouping very similar or exact order sets, while separating those that are completely different or have limited overlap in orders. The web interface also received strong support from the reviewers and the constructive criticism will help guide us in our development of a user-friendly and effective order set curation software.

Our future work will focus on continued development of the software to improve the usability based upon feedback from the reviewers. Additional functionality to enable users to create a separate curated database by selecting, removing, merging, and modifying existing order sets in the database will be considered, as well as features to aid the decision making. Lastly, a comprehensive inter-rater reliability study will be conducted on the full order set database with the goal of documenting output reliability. We believe the combination of NLP and clustering methodologies has great potential for the curation of Clinical Decision Support artifacts in ways that can reduce unnecessary variation within and among clinical sites in the VA network and other medical facilities. In so doing we believe we can improve care provided to Veterans.

Acknowledgements

This work has been supported in part by grants from NIH NLM T15LM012495, NIAAA R21AA026954, R33AA0226954 and NCATS UL1TR001412. This study was funded in part by the Department of Veterans Affairs.

References

[1] Kuperman GJ, Gibson RF. Computer physician order entry: benefits, costs, and issues. Annals of internal medicine. 2003 Jul 1;139(1):31-9.

[2] Wright A, Sittig DF, Carpenter JD, Krall MA, Pang JE, Middleton B. Order sets in computerized physician order entry systems: an analysis of seven sites. AMIA Annu Symp Proc. 2010 Nov 13;2010:892-6.

[3] Ozdas A, Speroff T, Waitman LR, Ozbolt J, Butler J, Miller RA. Integrating "best of care" protocols into clinicians' workflow via care provider order entry: impact on quality-of-care indicators for acute myocardial infarction. J Am Med Inform Assoc. 2006 Mar-Apr;13(2):188-96.

[4] Hulse NC, Lee J, Borgeson T. Visualization of Order Set Creation and Usage Patterns in Early Implementation Phases of an Electronic Health Record. AMIA Annu Symp Proc. 2017 Feb 10;2016:657-666.

[5] Muniga ET, Walroth TA, Washburn NC. The impact of changes to an electronic admission order set on prescribing and clinical outcomes in the intensive care unit. Applied clinical informatics. 2020 Jan;11(01):182-9.

[6] Brown SH, Lincoln MJ, Groen PJ, Kolodner RM. VistA--U.S. Department of Veterans Affairs national-scale HIS. *Int J Med Inf.* 2003;69(2–3):135–56.

[7] Resnick MP, LeHouillier F, Brown SH, Campbell KE, Montella D, Elkin PL. Automated Modeling of Clinical Narrative with High Definition Natural Language Processing Using Solor and Analysis Normal Form. Stud Health Technol Inform. 2021 Nov 18;287:89-93.

[8] Elkin PL, Mullin S, Mardekian J, Crowner C, Sakilay S, Sinha S, Brady G, Wright M, Nolen K, Trainer J, Koppel R, Schlegel D, Kaushik S, Zhao J, Song B, Anand E. Using Artificial Intelligence With Natural Language Processing to Combine Electronic Health Record's Structured and Free Text Data to Identify Nonvalvular Atrial Fibrillation to Decrease Strokes and Death: Evaluation and Case-Control Study. J Med Internet Res. 2021 Nov 9;23(11):e28946.

Challenges of Trustable AI and Added-Value on Health
B. Séroussi et al. (Eds.)

doi:10.3233/SHTI220503

The Prediction of Functional Outcome After Microsurgical Treatment of Unruptured Intracranial Aneurysm Based on Machine Learning

Timur ISHANKULOV[a], Fyodor GREBENEV[a], Uliya STRUNINA[a], Oleg SHEKHTMAN[a], Shalva ELIAVA[a] and Gleb DANILOV[a,1]
[a] *National Medical Research Center for Neurosurgery named after N.N. Burdenko, Moscow, Russian Federation*

Abstract. Our study aimed to create a machine learning model to predict patients' functional outcomes after microsurgical treatment of unruptured intracranial aneurysms (UIA). Data on 615 microsurgically treated patients with UIA were collected retrospectively from the Electronic Health Records at N.N. Burdenko Neurosurgery Center (Moscow, Russia). The dichotomized modified Rankin Scale (mRS) at the discharge was used as a target variable. Several machine learning models were utilized: a random forest upon decision trees (RF), logistic regression (LR), support vector machine (SVM). The best result with F1-score metric = 0.904 was produced by the SVM model with a label-encode method. The predictive modeling based on machine learning might be promising as a decision support tool in intracranial aneurysm surgery.

Keywords. Intracranial aneurysm, machine learning, classification, modified ranking scale, mRS

1. Introduction

Brain aneurysms occur in ~2.8% of the population and pose a severe threat to patients due to the risk of rupture and intracranial hemorrhage (1). Evidence-based management of unruptured intracranial aneurysms (UIA) might be grounded on the individual prognosis of aneurysm growth and rupture. A series of predictive modeling studies is known to address this issue. On the other hand, predicting treatment outcomes might be valued for balanced decision-making. Despite the apparent benefits of preventive treatment, surgery brings risks of disability (2.2%–10.9%) and death (0.0%–2.3%) (2). There is a lack of studies evaluating the performance of artificial intelligence in the latest task. V. Staartjes et al. (2020) demonstrated the power of machine learning to predict the outcome of unruptured aneurysm surgery in a pilot research (3). The validity of such an approach in various cohorts is a subject of rigorous investigation. Our study aimed to assess the quality of predicting the functional outcome after UIA microsurgical treatment using supervised machine learning on a single-center dataset.

[1] Corresponding Author, Gleb Danilov, N.N. Burdenko Neurosurgery Center, 4th Tverskaya-Yamskaya str. 16, Moscow 125047, Russian Federation; E-mail: glebda@yandex.ru.

2. Methods

Patients with unruptured intracranial aneurysms underwent microsurgery at N.N. Burdenko National Medical Research Center for Neurosurgery (Moscow, Russian Federation) in 2018-2021 were eligible for our study. We did not consider cases of consecutive UIA treatment (including endovascular) within one hospitalization, as well as aneurysms associated with other brain lesions (tumors, arteriovenous malformations, etc.). The study was approved by the Local Ethics Committee and Neurovascular Research Board at N.N. Burdenko Neurosurgery Center. Informed consent was obtained for all patients.

We collected data retrospectively from the Electronic Health Records. A set of categorical and numeric variables reflecting various patients' characteristics (age, gender, history of past illness, disease severity, neurological status, neuroimaging, intraoperative data) was first investigated via exploratory data analysis and preprocessed with feature engineering techniques. Table 1 demonstrates the final feature space we selected to use in models. Aneurysm size was evaluated via preoperative radiological data (computed tomography, magnetic resonance, or digital subtraction angiography) and matched with surgical reports and intraoperative video recordings. Thus, nine predictors related directly to patients, nine – to aneurysms, and six variables - to surgery were chosen (Table 1).

Table 1. Features related to patients, UIA and surgery exploited in predictive modeling.

	Patient's characteristics	Aneurysm features	Surgical parameters
1	Sex	Localization	Simultaneous surgery for multiple aneurysms
2	Age	Shape	Intraoperative bypass
3	American Society of Anesthesiologists (ASA) physical status	Wind neck (aneurysm neck equal or bigger than parent artery diameter)	Blood clotting disorder due to working anticoagulant/antiplatelet therapy and/or coagulopathy during the surgery
4	mRS before surgery	Size	Retrograde suction decompression or direct blood aspiration
5	History of stroke	Diverticula	
6	Number of functioning UIAs	Vessels/Nerve structures involvement	Neurosurgeon's experience (frequently or rarely operating)
7	Symptomatic type of UIA treated	Calcification/atherosclerotic lesion of aneurysm or parent artery	
8	History of anticoagulants/antiplatelet agents uptake or coagulopathy	Intraluminal thrombosis	
9	History of other intracranial aneurysms treatment	Repeated UIA treatment	

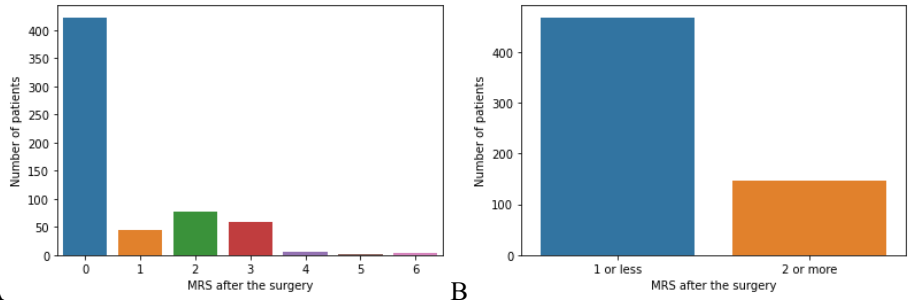
The majority of non-binary quantitative and multilevel categorical variables were collapsed into categorical to reduce the number of within-variable strata.

A modified Rankin Scale (mRS) assessment on the day of discharge was chosen as the basis for the target variable construction. mRS after the surgery varied from 0 (no

symptoms) to 6 (death) and showed a pronounced imbalance which we tried to address primarily with dichotomization (Figure 1B). The binary target variable took a value of 0 if postoperative mRS was 1 or less and a value of 1 if it was equal to 2 or greater (Figure 1B).

Several machine learning models were used to predict the target variable: a random forest over decision trees (RF), logistic regression (LR), support vector machine (SVM). We also applied two labeling methods: simple label and one-hot encoding for each machine learning model. Thus, a total of 6 models were introduced.

Each test was performed after the data were randomly sampled into training (80%) and testing (20%) subsets with stratification. The model was trained on a training subset; 5-fold cross-validation (CV) was utilized to evaluate the model's quality before the final testing. Each machine learning model was tested 300 times with stratified resampling (1800 tests in total).



A B

Figure 1. A - distribution of original mRS upon admission; B - distribution of binary target variable in the preprocessed dataset.

We used standard metrics to evaluate the test results: accuracy on validation samples within the cross-validation (*CV*), *Accuracy*, *Precision*, *Recall* and *F1-score* on testing samples. The results for each machine learning model were averaged across all metrics to account for the random data split and reduce the margin of error.

The exploratory data analysis was performed within the R programming environment (version 4.0.3) in RStudio Server IDE (version 1.3.1093). The modelling was done using the Python programming language (version 3.7) with the *pandas, numpy* and *sklearn* libraries in Jupyter Notebook.

3. Results

A total of 615 patients were enrolled in the study. In 31 cases, aneurysm size could not be retrieved and was denoted as "not specified". The "zeros" class has 468 points, and "ones" included 147 values (Figure 1, B).

The results of our classification experiments are presented in Table 2 in descending order of F1-score.

Table 2. UIA microsurgery outcome prediction quality with three machine learning models and two labeling methods. The 95% confidence intervals for F1 (in square brackets) were obtained via bootstrapping.

Model	Encoding	CV	Precision	Recall	Accuracy	F1-score
SVM	Label	0.922	0.951	0.878	0.925	0.904 [0.901, 0.907]
LR	Label	0.922	0.949	0.878	0.924	0.903 [0.900, 0.906]
SVM	One-hot	0.922	0.949	0.876	0.923	0.902 [0.899, 0.905]
LR	One-hot	0.919	0.941	0.872	0.918	0.896 [0.893, 0.899]
RF	One-hot	0.913	0.927	0.872	0.914	0.892 [0.888, 0.895]
RF	Label	0.911	0.926	0.871	0.913	0.891 [0.888, 0.895]

The best result in terms of the F1-score metric was 0.904, achieved with the SVM model and label-encoding method. However, the other models demonstrated close results. The results support the hypothesis that the binary outcomes of UIA surgery are pretty well separable using the proposed feature set despite imbalance.

Transformation of variables into categorical type with a reduced number of strata led to a better machine learning performance. Our resampling approach enabled calculations with a low margin of error (< 0.005). Label encoding exerted a minor influence on the results.

4. Discussion

As the availability and quality of neuroimaging improve, the number of patients diagnosed with UIA increases. In this regard, the number of operations on UIAs is also growing. For example, the number of operations for brain aneurysms in the Russian Federation increased from 1278 (in 2007) up to 7281 (in 2017), with approximately 40% of surgery performed for UIA [8]. In 60.0%-62.7% of cases, microsurgical clipping was preferable for UIA treatment in our country. A successful operation saves the patient from aneurysm rupture. On the other hand, any surgical treatment carries the risks of disability and mortality.

The most rational approach in UIA management is to compare the probability of aneurysm growth and rupture during observation with the hazards of surgical treatment. Despite the studies UCAS (4), PHASES (5,6), and ELAPSS (7,8) could give a numerical answer to the question: “What are the risks of aneurysm growth and rupture in observation?”, the UIATS study (9,10) could not answer the similar question regarding surgical interventions.

A way to cope with the prediction task is to use machine learning, which was successfully done by V. Staartjes et al. in 2020 (3). The authors demonstrated good models performance (AUC of 0.63–0.77 and the accuracy of 0.78–0.91, respectively). That work differs from our approach by less amount of input data, non-accounting of such intraoperative features as thrombextraction, retrograde suction decompression, non-inclusion of patients under 17 years of age, as well as the use of UIATS scores as one of the parameters, which could be partly subjective, and also implied not only microsurgical but endovascular treatment as well. We took that into account while collecting our database.

The limitations of our study were related to a single-center proprietary dataset, a limited number of models tested, initial sample imbalance, no additional sampling to

address the imbalance, restricted feature space, no prospective model validation. Future work should be aimed at overpassing these shortcomings.

5. Conclusions

In this pilot study, good quality of functional outcome prediction after microsurgical treatment of UIA was demonstrated using traditional (shallow) machine learning methods. The predictive modeling based on machine learning might be promising as a decision support tool concerning intracranial aneurysm surgery.

References

[1] Vlak MHM, Algra A, Brandenburg R, Rinkel GJE. Prevalence of unruptured intracranial aneurysms, with emphasis on sex, age, comorbidity, country, and time period: a systematic review and meta-analysis. Lancet Neurol. 2011 Jul 1;10(7):626–36.

[2] Algra AM, Lindgren A, Vergouwen MDI, Greving JP, Van Der Schaaf IC, Van Doormaal TPC, et al. Procedural Clinical Complications, Case-Fatality Risks, and Risk Factors in Endovascular and Neurosurgical Treatment of Unruptured Intracranial Aneurysms: A Systematic Review and Meta-analysis. JAMA Neurol [Internet]. 2019 Mar 1 [cited 2022 Mar 28];76(3):282–93. Available from: https://pubmed.ncbi.nlm.nih.gov/30592482/

[3] Staartjes VE, Sebök M, Blum PG, Serra C, Germans MR, Krayenbühl N, et al. Development of machine learning-based preoperative predictive analytics for unruptured intracranial aneurysm surgery: a pilot study. Acta Neurochir (Wien) [Internet]. 2020 Nov 1 [cited 2022 Mar 28];162(11):2759–65. Available from: https://link.springer.com/article/10.1007/s00701-020-04355-0

[4] Morita A, Kirino T, Hashi K, Aoki N, Fukuhara S, Hashimoto N, et al. The natural course of unruptured cerebral aneurysms in a Japanese cohort. N Engl J Med [Internet]. 2012 Jun 28 [cited 2022 Mar 28];366(26):2474–82. Available from: https://pubmed.ncbi.nlm.nih.gov/22738097/

[5] Greving JP, Wermer MJH, Brown RD, Morita A, Juvela S, Yonekura M, et al. Development of the PHASES score for prediction of risk of rupture of intracranial aneurysms: a pooled analysis of six prospective cohort studies. Lancet Neurol [Internet]. 2014 Jan [cited 2022 Mar 28];13(1):59–66. Available from: https://pubmed.ncbi.nlm.nih.gov/24290159/

[6] Backes D, Vergouwen MDI, Tiel Groenestege AT, Bor ASE, Velthuis BK, Greving JP, et al. PHASES Score for Prediction of Intracranial Aneurysm Growth. Stroke [Internet]. 2015 May 20 [cited 2022 Mar 28];46(5):1221–6. Available from: https://www.ahajournals.org/doi/abs/10.1161/strokeaha.114.008198

[7] Backes D, Rinkel GJE, Greving JP, Velthuis BK, Murayama Y, Takao H, et al. ELAPSS score for prediction of risk of growth of unruptured intracranial aneurysms. Neurology [Internet]. 2017 Apr 25 [cited 2022 Mar 28];88(17):1600–6. Available from: https://pubmed.ncbi.nlm.nih.gov/28363976/

[8] van Kammen MS, Greving JP, Kuroda S, Kashiwazaki D, Morita A, Shiokawa Y, et al. External Validation of the ELAPSS Score for Prediction of Unruptured Intracranial Aneurysm Growth Risk. J Stroke [Internet]. 2019 Sep 1 [cited 2022 Mar 28];21(3):340. Available from: /pmc/articles/PMC6780020/

[9] Etminan N, Beseoglu K, Barrow DL, Bederson J, Brown RD, Connolly ES, et al. Multidisciplinary consensus on assessment of unruptured intracranial aneurysms: Proposal of an international research group. Stroke [Internet]. 2014 [cited 2022 Mar 28];45(5):1523–30. Available from: https://www.ahajournals.org/doi/abs/10.1161/strokeaha.114.004519

[10] Etminan N, Brown RD, Beseoglu K, Juvela S, Raymond J, Morita A, et al. The unruptured intracranial aneurysm treatment score: A multidisciplinary consensus. Neurology [Internet]. 2015 Sep 8 [cited 2022 Mar 28];85(10):881. Available from: /pmc/articles/PMC4560059/

Challenges of Trustable AI and Added-Value on Health
B. Séroussi et al. (Eds.)

doi:10.3233/SHTI220504

Clinical Decision Support System for PIM in Elderly Patients: Implementation and Initial Evaluation in Ambulatory Care

Eliana FRUTOS[a,1] , Martin KAKAZU[a], Matias TAJERIAN[a], Alejandro GAIERA[a]
Luciana RUBIN[a], Carlos OTERO[a] and Daniel LUNA[a]
[a] *Health Informatics Department, Hospital Italiano de Buenos Aires, Argentina*

Abstract. The high prevalence of PIMs in elderly is a major healthcare concern and indicates the need for medication monitoring systems. Most PIM CDSS have shown positive effects respecting PIM prescription but these results were more consistently in hospital settings compared with ambulatory care. We describe the post-implementation evaluation of a PIM CDSS for general practitioners (GP) in the ambulatory setting and explore GP interactions with the PIM alerts. The CDSS generated 3218 unique alerts and involved 2863 elderly patients. Benzodiazepines was the drug with the most alerts triggered. Only 129 (4 %) were opened by GP during patient appointments. We need to develop an understanding of how alerts should be designed and display information to support the workflow of general practitioners. Pos-implementation evaluations are the key of CDSS improvements.

Keywords. Clinical Decision Support systems, Potentially Inappropriate Medication, Elderly, Electronic Health Records.

1. Introduction

The world's population is rapidly aging and is projected to double nearly 1.5 billion in 2050 [1]. The increase in multimorbidity results in a high use of medications in the elderly where potentially inappropriate medications (PIM) become more prevalent [2]. PIM may be defined as medications which result in a significant risk of adverse health outcomes, are associated with adverse drug events, morbidity, functional decline, and mortality [3].

The high prevalence of PIM is a major healthcare concern and indicates the need for medication monitoring systems. To address this issue, many guidelines have been developed. One example is Beers Criteria, an important tool for physicians in medication management in elderly [4].

Reducing PIM has been challenging and many interventions were implemented in this matter (educational assessment, restrictive prescriptions, etc). The benefits of Clinical Decision Support Systems (CDSS) regarding medical error and health outcomes are widely known [5]. Most PIM CDSS have shown positive effects respecting PIM

[1] Corresponding Author, Eliana Frutos, Health Informatics Department, Hospital Italiano de Buenos Aires, Tte. Gral. Juan D. Perón 4190, C1199ABB Buenos Aires, Argentina; E-mail: eliana.frutos@hospitalitaliano.org.ar.

prescription, but these results were more consistently in hospital settings compared with ambulatory care [6]. This gap between settings reveals the need to investigate ambulatory care.

The goal of this study is to describe the post-implementation evaluation of a PIM CDSS for general practitioners (GP) in the ambulatory setting and explore GP interactions with the PIM alerts.

2. Methods

2.1. Setting

Our study took place at Hospital Italiano de Buenos Aires (HIBA), a community-based tertiary care hospital located in Buenos Aires, Argentina. HIBA also has outpatient clinics in its network. It is a HIMSS Level 7+ organization with an in-house developed health information system which includes clinical and administrative data. It features a web-based, problem-oriented EHR; a terminology server referenced to SNOMED CT; and an integrated personal health record (PHR). HIBA has its health insurance called Plan de Salud (PS) with over 150,000 affiliates, 23 % of them being elderly.

2.2. Design and data collection

We conducted a cross sectional study between August and December 2021. We analyzed all PIM CDSS alerts triggered in EHR during the study period.

Regarding data collection, we used secondary databases with clinical and administrative data sets containing different variables of interest related with PIM CDSS: PIM type, patient's sex and age and GP interactions with the alert (click-through rate).

2.3. PIM CDSS

A PIM Clinical decision tool was designed by an interdisciplinary team of health informatics specialists and General Practitioners to address the need for identification of elderly outpatients who purchased PIM at the HIBA pharmacy.

A highlighted icon in the header of the patient's EHR shows that the clinical alert has been triggered, and clicking it displays the alert in detail (Figure 1).

The alert triggers when the following criteria are met: 1) Patient is over 64 years old, 2) PS affiliated, 3) Patient has a GP assigned, 4) PIM purchase in a period no longer than 92 days prior, 5) Scheduled appointment with GP (face-to-face or teleconsultation). Patients without appointments or hospitalized were excluded.

Three sets of PIM were defined based on the Beers criteria and on data collected in previous studies in HIBA. The three most common drug classes prescribed in the elderly were selected: Benzodiazepines (BZDs), Proton Pump Inhibitors (PPIs) and Non-steroidal anti-inflammatories (NSAIDs) [7].

The CDSS alert provides information about PIM purchase (date and PIM type), includes links to educational information as Beers Criteria 2019, and suggests prescription revision. Also includes clinical algorithms regarding PIM to guide GP. It has some features like the 'Go to Prescription module' button which automatically

redirects to the EHR prescription module, “Suggestions” and “Don’t show again” which deactivates the alert for a period of 92 days.

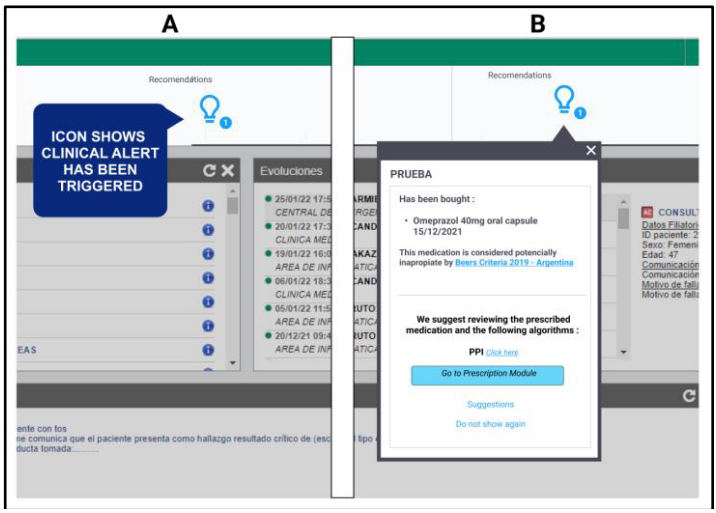


Figure 1. PIM CDSS Alert

Fast Healthcare Interoperability Resources (FHIR), CDS-Hooks and SNOMED CT standards were used for the design and development of this tool.

2.4. Ethical considerations

The study was performed in full agreement with current national and international ethical regulations. Was approved by the institutional ethics committee (CEPI # 6174).

3. Results

The CDSS generated 3218 unique alerts during the study period. It involved 2863 elderly patients with a mean age of 78 (DS 7.6). Female sex was predominant with 74.85 % (2143). Almost half of the patients triggered the alert more than twice for different drug purchases during this period.

Regarding PIM, 52.37 % (2280) belonged to BZDs, 40.20 % (1750) PPIs and only 7.05 % (307) were NSAIDs. Analyzing the BZDs group, it was composed mostly of Clonazepam in 35% (820) and Alprazolam in 33.64% (768). The rest was composed of a variable proportion of Lorazepam, Bromazepam and Midazolam. Respecting PPIs, Omeprazol and Esomeprazol were the most frequent (48% , 23.45%). Finally, half of NSAIDs were Diclofenac (51.79 %) followed by Ibuprofen (24.75%) and Ketorolac (28.89%).

Among 3218 alerts triggered and available in the EHR as a highlighted lightbulb Icon, only 129 (4 %) were opened by GP during patient appointments. Exploring actions performed in the alert by users, suggestions and prescription module’s buttons had a click-through rate (CTR) of 8.95 % and 7.25% respectively. We found a low CRT in

educational links to see information about PPIs and NSAIDs. The actions most frequently taken were opening and closing the alert without any interaction. (Table 1).

Table 1. General Practitioners clicks interaction in alert features

Action Description	Alerts opened (129) Total Clicks = 151	Alerts Opened (75) In - person visit (Clicks = 117)	Alerts Opened (54) Teleconsultation (Clicks = 34)
Go to prescription module button	5.96 % (9)	6.83% (8)	2.94% (1)
Suggestion button	7.28% (11)	3.41% (4)	8.82% (3)
Do not show again button	3.31 % (5)	3.41% (4)	2.94% (1)
Educational links about PIM			
PPIs information link	3.97 % (6)	3.41% (4)	5.88% (2)
NSAIDs information link	0.66% (1)	0.85% (1)	-

4. Discussion

In this study, we attempted to evaluate a CDSS which identifies elderly patients with PIM purchased, with the aim of suggesting to general practitioners to do an evaluation of PIM prescription.

Potentially inappropriate medications among older patients have a high prevalence reported by Storms et al. [8]. The amount of alerts triggered in our findings may be an indicator about how much elderly population consumed PIM, regardless of the potential adverse effects [9]. Regarding PIM, our results showed a low percent of NSAIDs compared to what is reported in literature [10]. This situation can be explained by the fact that in Argentina this type of drug is very accessible and it is sold in many shops without prescription.

General practitioners overrode 96% of PIM alerts. Very few GP clicked on educational links and followed the alert's suggestion. This finding is consistent with high override rates reported in the literature [11]. Many factors could be involved in this override rate like alert design and care setting. The outpatient care workflow is non-identical to a hospitalized environment in which more of the CDSS evaluation studies focus [12]. Qualitative research is needed to better understand how physicians interact with decision support at the point of care and why they ignored it. We believe that override rate monitoring should be a standard practice if a CDSS is implemented.

This study has some limitations. Our CDSS is based only on a single clinical guideline (Beers Criteria) and elderly patients are multi-morbid patients who require the integration of multiple clínical practice guidelines for a better approach [13]. Also, the older adults who triggered the alert may not be representative of the general older adult population in Argentina. Respecting PIM, we evaluated only 3 sets of medicines (BZDs, PPIs and NSAIDs) which limited the fully comprehension of PIM purchased in elderly.

As a future line of work, we will redesign alert's features and visualization display with the intention to explore the effect in general practitioners's response.

5. Conclusion

PIM still are a major health concern in elderly as our study showed in the amount of alerts triggered. More research is necessary, applying a usability approach, to explore general practitioners needs and patterns of use for a full adoption of PIM CDSS.

References

[1] United Nations, Department of Economic and Social Affairs, Population Division (2019). World Population Ageing 2019: Highlights (ST/ESA/SER.A/430)

[2] Storms H, Marquet K, Aertgeerts B, Claes N. Prevalence of inappropriate medication use in residential long-term care facilities for the elderly: A systematic review. Eur J Gen Pract. 2017 Dec;23(1):69-77. doi: 10.1080/13814788.2017.1288211. PMID: 28271916; PMCID: PMC5774291.

[3] Cahir C, Bennett K, Teljeur C, Fahey T. Potentially inappropriate prescribing and adverse health outcomes in community dwelling older patients. Br J Clin Pharmacol. 2014 Jan;77(1):201-10. doi: 10.1111/bcp.12161. PMID: 23711082; PMCID: PMC3895361.

[4] By the 2019 American Geriatrics Society Beers Criteria® Update Expert Panel. American Geriatrics Society 2019 Updated AGS Beers Criteria® for Potentially Inappropriate Medication Use in Older Adults. J Am Geriatr Soc. 2019 Apr;67(4):674-694. doi: 10.1111/jgs.15767. Epub 2019 Jan 29. PMID: 30693946.

[5] Beeler PE, Bates DW, Hug BL. Clinical decision support systems. Swiss Med Wkly. 2014 Dec 23;144:w14073. doi: 10.4414/smw.2014.14073. PMID: 25668157.

[6] Ghibelli S, Marengoni A, Djade CD, Nobili A, Tettamanti M, Franchi C, Caccia S, Giovarruscio F, Remuzzi A, Pasina L. Prevention of inappropriate prescribing in hospitalized older patients using a computerized prescription support system (INTERcheck(®)). Drugs Aging. 2013 Oct;30(10):821-8. doi: 10.1007/s40266-013-0109-5. PMID: 23943248.

[7] Schapira M, Calabró P, Montero-Odasso M, Osman A, Guajardo ME, Martínez B, Pollán J, Cámera L, Sassano M, Perman G. A multifactorial intervention to lower potentially inappropriate medication use in older adults in Argentina. Aging Clin Exp Res. 2021 Dec;33(12):3313-3320. doi: 10.1007/s40520-020-01582-4. Epub 2020 May 9. PMID: 32388838.

[8] Storms H, Marquet K, Aertgeerts B, Claes N. Prevalence of inappropriate medication use in residential long-term care facilities for the elderly: A systematic review. Eur J Gen Pract. 2017 Dec;23(1):69-77. doi: 10.1080/13814788.2017.1288211. PMID: 28271916; PMCID: PMC5774291.

[9] Dedhiya SD, Hancock E, Craig BA, Doebbeling CC, Thomas J 3rd. Incident use and outcomes associated with potentially inappropriate medication use in older adults. Am J Geriatr Pharmacother. 2010 Dec;8(6):562-70. doi: 10.1016/S1543-5946(10)80005-4. PMID: 21356505.

[10] Davidoff AJ, Miller GE, Sarpong EM, Yang E, Brandt N, Fick DM. Prevalence of potentially inappropriate medication use in older adults using the 2012 Beers criteria. Journal of the American Geriatrics Society. 2015 Mar;63(3):486-500.

[11] Slight SP, Seger DL, Nanji KC, Cho I, Maniam N, Dykes PC, Bates DW. Are we heeding the warning signs? Examining providers' overrides of computerized drug-drug interaction alerts in primary care. PLoS One. 2013 Dec 26;8(12):e85071. doi: 10.1371/journal.pone.0085071. PMID: 24386447; PMCID: PMC3873469.

[12] Onatade R, Auyeung V, Scutt G, Fernando J. Potentially inappropriate prescribing in patients on admission and discharge from an older peoples' unit of an acute UK hospital. Drugs Aging. 2013 Sep;30(9):729-37. doi: 10.1007/s40266-013-0097-5. PMID: 23780641..

[13] Michalowski M, Rao M, Wilk S, Michalowski W, Carrier M. MitPlan 2.0: Enhanced Support for Multi-morbid Patient Management Using Planning. In International Conference on Artificial Intelligence in Medicine 2021 Jun 15 (pp. 276-286). Springer, Cham..

Challenges of Trustable AI and Added-Value on Health
B. Séroussi et al. (Eds.)

doi:10.3233/SHTI220505

Towards the Improvement of Clinical Guidelines Based on Real World Data

Elisa HENKE[a,1], Ines REINECKE[a], Michele ZOCH[a], Martin SEDLMAYR[a], and Franziska BATHELT[a]
[a] *Institute for Medical Informatics and Biometry, Carl Gustav Carus Faculty of Medicine, Technische Universität Dresden, Dresden, Germany*

Abstract. Computerized clinical guidelines (CCG) are effective instruments for standardizing, monitoring and optimizing medical treatment processes. Nevertheless, due to barriers in flexibility, transferability and acceptance, the widespread use of CCG in clinical practice is not yet common. To overcome those issues, we present a concept on how to use real world data to evaluate CCG and to recommend improvements. As a first result, we defined an algorithm to extract treatment processes based on the standardized Observational Medical Outcomes Partnership (OMOP) Common Data Model as well as their visualization using the graphical modeling language Business Process Model and Notation (BPMN).

Keywords. Business Process Model and Notation, OMOP CDM, computerized clinical guidelines, interoperability

1. Introduction

Computerized clinical guidelines (CCG) are effective instruments for standardizing, monitoring and significantly improving processes in healthcare [1][2]. Nevertheless, barriers exist that prevent the adaption in clinical practice. Studies agree the most relevant reasons for those issues are: the lack of flexibility, transferability, acceptance and trust, resulting from a lack of evidence based on real world data (RWD) [3][4][5]. To address these gaps, Jin et.al. suggest to compact information through visualizations and to link CCG with electronic health record (EHR)-data [5].

To support technical transferability and comparability it is beneficial to use a standardized common data model (CDM) including standardized vocabularies and terminologies that ensure semantic interoperability across countries and healthcare fields. The Observational Health Data Science and Informatics (OHDSI) [6] provides the Observational Medical Outcomes Partnership (OMOP) CDM that gained significant relevance in the last years for research on RWD [7]. For comprehensive process visualization Business Process Model and Notation (BPMN) is considered as valuable asset in the healthcare section [8]. In this paper, we present a concept based on OMOP CDM and BPMN on how to use RWD to evaluate CCG and to recommend

[1] Corresponding Author, Elisa Henke, Institute for Medical Informatics and Biometry, Carl Gustav Carus Faculty of Medicine, Technische Universität Dresden, Dresden, Germany; E-mail: Elisa.Henke@tu-dresden.de.

improvements. We show preliminary results addressing the extraction of treatment processes for single patients as BPMN models based on OMOP CDM data.

2. Methods

2.1 Concept

Our concept (Figure 1) is based on EHR data stored in the OMOP CDM, which implies a standardization of data and allows comparability independent from the location of the hospital. Leveraging the OHDSI software stack allows researchers to define and share cohort definitions among sites. A cohort definition groups different patients with similar characteristics such as the treatment process. This information stored on the OMOP CDM database level can be further analyzed regarding patterns, i.e. common parts in the treatment plans, that are promising for positive outcomes. With that, a common clinical treatment process could be synthesized. This common treatment process is evaluated based on retrospective data with positive and negative outcomes. If it is hereby indicated that the common treatment process may increase the possibility of positive outcomes, it can be utilized to evaluate an associated CCG by finding contradictions, differences or overlaps, which results in recommendation for CCG related improvements. If synthesis of a common process is not possible due to too many differences in local treatment plans, it may be possible to synthesize several common treatment plans and assesses their possibilities for positive outcomes against each other. If that is not possible either, the treatment plan elements can be prioritized.

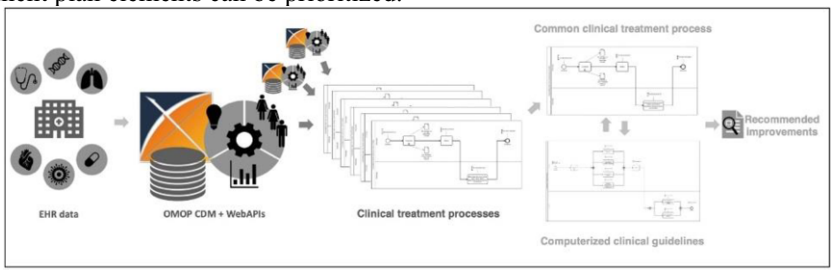


Figure 1: From EHR data to recommended CCG improvements – a concept (own illustration)

2.2 Requirements for Automatic Extraction of Treatment Processes from OMOP CDM

According to the data stored in OMOP CDM an algorithm to extract treatment processes as models in BPMN version 2.0 was defined [9]. In an interdisciplinary team, we have manually analyzed existing BPMN models of clinical treatment processes [10][11] and agreed on criteria for a BPMN model to be sufficient, when (a) for a single medical case, (b) treatment relevant parameters, (c) the chronological order of events, (d) parallel events and (e) involved actors are shown in the visualization.

2.3 Access the process relevant data within the OMOP CDM

Process relevant data elements required in OMOP CDM were identified based on the analysis of the guidelines for stroke treatments. These data elements are stored in

following tables within OMOP CDM v5.3.1. according to the implementation of Gruhl et.al [12]: VISIT_OCCURRENCE, MEASUREMENT, CONDITION_OCCURRENCE, PROCEDURE_OCCURRENCE, and VISIT_DETAIL, CARE_SITE and CONCEPT.

2.4 Automatically generate BPMN models

First, we designed an algorithm that maps the process relevant OMOP CDM data elements identified in section 2.3 to BPMN elements. Second, we implemented the algorithm in Java as a prototype and tested it with anonymized data based on the core data set of the Medical Informatics Initiative Germany [13]. For visualization purpose the automatically generated treatment processes were imported into the software Camunda Modeler (Version 4.0.0) [14] via its interface. Our implementation automatically generates a .bpmn file out of OMOP CDM data that meets the interface specification.

3. Results

3.1 Link relevant data to BPMN model element

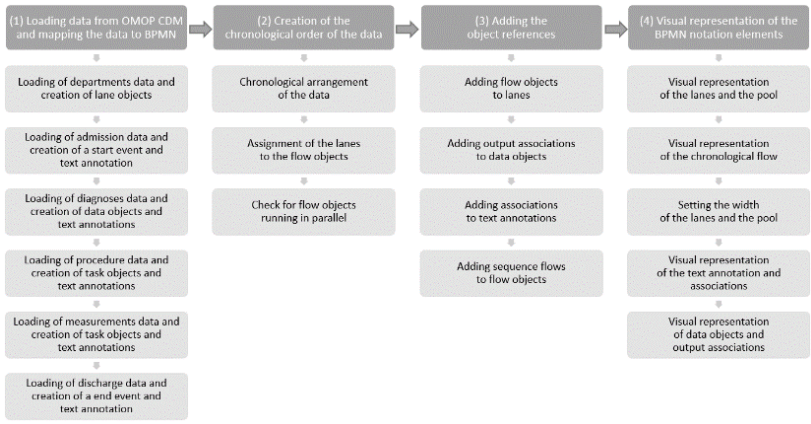


Figure 2: Algorithm to automatically model treatment processes with BPMN by extracting data from OMOP CDM (own illustration)

The necessary data elements (see section 2.3) are mapped to BPMN elements (Figure 2 – step 1). The complete mapping table is accessible through the supplemental files[2]. In step 2 data is put into a chronological order based on the comparison of timestamps and followed by the arrangement and assignment of the data to corresponding "*lanes*". As different events may have the same timestamps, the data is also checked for concurrency. If parallel events occur, two *"parallel gateways"* are used for branching and merging the process flow. In the third step object references are assigned, that specify which BPMN notation elements are connected to each other. Step 4 comprises the complete visual

[2] https://caruscloud.uniklinikum-dresden.de/index.php/s/8cNfm8kpwCdJ63e

representation of the whole treatment process including automatic layouting of the constructed elements.

3.2 Automatically generate BPMN models

The method defined in section 2.4 was applied on five medical cases of different patients. It resulted in BPMN models as roughly shown by example in Figure 3. The exact models can be found in the supplemental files[2]. Each resulting BPMN model has confirmed being correct according to original data by at least two persons.

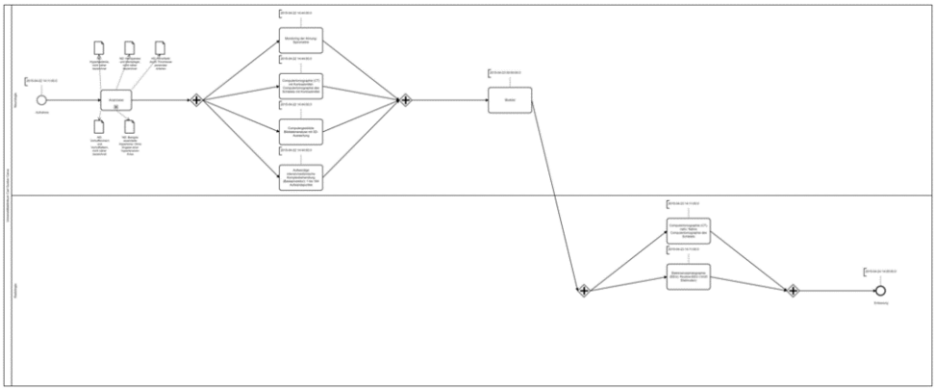


Figure 3: Automatically created BPMN model (own illustration; tool: Camunda Modeler)

4 Discussion & Conclusion

To the best of our knowledge, there is no approach that uses data stored in OMOP CDM to gain evidence for the accuracy of CCGs [7]. With the presented extraction, a first promising step to realize and evaluate the concept to improve CCGs based on RWD stored in OMOP CDM has been made. This enables us to automatically extract patient treatment processes from OMOP based data in more detail than existing approaches that are able to visualize a patient related clinical report including a summary (amount) of observations, procedures, diagnoses etc. [15], to find similarities in medical treatments like diabetes, hypertension and depressions [16] or for pediatric epilepsy [17]. The OHDSI applications ATLAS “Cohort Pathways” and “Profiles” [6], visualize a summary of medical treatments for specific conditions over a patient cohort as sunburst plot and support a visualization for events over a time period, representing clinical events within different domains (e.g. observations, measurements, etc.), respectively. However, a representation of the treatment process for a single visit including involved actors is not realizable with these approaches. Our approach can represent the treatment process as BPMN model and meets the initially stated criteria (section 2.2). With that, we addressed the strategies stated by Jin et.al. [5] namely information visualization and linkage to EHR data. Nonetheless, our approach is a prototype that has been evaluated only on five medical cases and is currently limited to a restricted set of parameters, due to the initial focus on the diagnosis stroke. Despite these required optimizations, we will work on the synthesis of an optimized treatment process from the set of treatment processes, which

comes with immense challenges and the comparison to specific guidelines. For both cases, we aim to evaluate the options delivered by BPMN and the comparison of different models [18].

Declaration

Conflict of Interest: The authors declare that there is no conflict of interest.

Author contributions: EH and FB: conception of the work. FB defined the concept, while EH designed and implemented the OMOP CDM to BPMN algorithm. All authors contributed substantial ideas and participated in editing and revising of the manuscript. All authors approved the manuscript in the submitted version and take responsibility for the scientific integrity of the work.

References

[1] Damiani G, Pinnarelli L, Colosimo SC, et al. The effectiveness of computerized clinical guidelines in the process of care: a systematic review. BMC Health Serv Res. 2010;10:2.
[2] Neame MT, Chacko J, Surace AE, et al. A systematic review of the effects of implementing clinical pathways supported by health information technologies. Journal of the American Medical Informatics Association. 2019;26:356–363.
[3] Trivedi MH, Kern JK, Marcee A, et al. Development and implementation of computerized clinical guidelines: barriers and solutions. Methods Inf Med. 2002;41:435–442.
[4] Forsner T, Hansson J, Brommels M, et al. Implementing clinical guidelines in psychiatry: a qualitative study of perceived facilitators and barriers. BMC Psychiatry. 2010;10:8.
[5] Jin Y, Li Z, Han F, et al. Barriers and enablers for the implementation of clinical practice guidelines in China: a mixed-method study. BMJ Open. 2019;9:e026328.
[6] Observational Health Data Sciences and Informatics (OHDSI). The Book of OHDSI [Internet]. 2021 [cited 2022 Jan 13]. Available from: https://ohdsi.github.io/TheBookOfOhdsi/.
[7] Reinecke I, Zoch M, Reich C, et al. The Usage of OHDSI OMOP - A Scoping Review. Stud Health Technol Inform. 2021;283:95–103.
[8] Stefanelli M. Knowledge and process management in health care organizations. Methods Inf Med. 2004;43:525–535.
[9] Object Management Group. Business Process Model and Notation (BPMN): Version 2.0 [Internet]. 2011 [cited 2022 Jan 13]. Available from: https://www.omg.org/spec/BPMN/2.0/PDF.
[10] Duske, Müller R, Prüfer R, et al. A BPMN Model of the Charite Stroke Treatment Process. Informatik-Berichte 244. 2014.
[11] Sooter LJ, Hasley S, Lario R, et al. Modeling a Clinical Pathway for Contraception. Appl Clin Inform. 2019;10(5):935-943.
[12] Gruhl M, Reinecke I, Sedlmayr M. Specification and Distribution of Vocabularies Among Consortial Partners. Stud Health Technol Inform. 2020;270:1393–1394.
[13] Medical Informatics Initiative Germany (MI-I). The Medical Informatics Initiative's core data set [Internet]. [cited 2022 Jan 13]. Available from: https://www.medizininformatik-initiative.de/en/medical-informatics-initiatives-core-data-set.
[14] Camunda Modeler [Internet]. [cited 2022 Jan 13]. Available from: https://camunda.com/products/camunda-platform/modeler/.
[15] Glicksberg BS, Oskotsky B, Thangaraj PM, et al. PatientExploreR: an extensible application for dynamic visualization of patient clinical history from electronic health records in the OMOP common data model. Wren J, editor. Bioinformatics. 2019;35:4515–4518.
[16] Zhang X, Wang L, Miao S, et al. Analysis of treatment pathways for three chronic diseases using OMOP CDM. J Med Syst. 2018;42:260.
[17] Kim H, Yoo S, Jeon Y, et al. Characterization of Anti-seizure Medication Treatment Pathways in Pediatric Epilepsy Using the Electronic Health Record-Based Common Data Model. Front Neurol. 2020;11:409.
[18] Kalenkova A, Polyvyanyy A, Rosa ML. Structural and Behavioral Biases in Process Comparison Using Models and Logs. In: Ghose A, Horkoff J, Silva Souza VE, et al., editors. Conceptual Modeling [Internet]. Cham: Springer International Publishing; 2021 [cited 2022 Jan 18]. p. 62–73. Available from: https://link.springer.com/10.1007/978-3-030-89022-3_6.

Challenges of Trustable AI and Added-Value on Health
B. Séroussi et al. (Eds.)

doi:10.3233/SHTI220506

Constructive Fuzzy Cognitive Map for Depression Severity Estimation

Georgia SOVATZIDI[a], Michael VASILAKAKIS, and Dimitris K. IAKOVIDIS[a,1]
[a] *Dept of Computer Science and Biomedical Informatics, Univ. of Thessaly, Greece*

Abstract. Depression is a common and serious medical disorder that negatively affects the mood and the emotions of people, especially adolescents. In this paper, a novel framework for automatically creating Fuzzy Cognitive Maps (FCMs) is proposed. It is applied for the estimation of the severity of depression among adolescents, based on their electroencephalogram (EEG). The introduced Constructive FCM (CFCM) utilizes features extracted by a Constructive Fuzzy Representation Model (CFRM), which conduces to detect in a more intuitive way the cause-and-effect relationships between the brain activity and depression. CFCM contributes to limiting the participation of experts, and the manual interventions in the traditional construction of FCMs, it provides an embedded mechanism for dimensionality reduction, and it constitutes an inherently interpretable approach to decision making, while being uncertainty-aware and simple to implement. The results of the experiments, using a recent publicly available dataset, demonstrate the effectiveness of the proposed framework and highlight its advantages.

Keywords. Fuzzy Cognitive Map, Fuzzy logic, Artificial Intelligence, Electroencephalogram (EEG), Interpretability, Depression.

1. Introduction

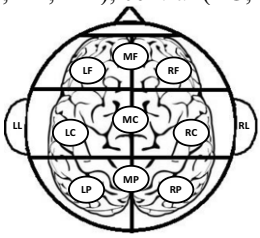
Clinical depression is a common mental disorder affecting more than 264 million people worldwide, according to the World Health Organization (1). The incidence of depression, like other related mood disorders, increases dramatically during the adolescence (2). Each adolescent may experience this disorder with different symptoms, such as emotional and cognitive signs, *e.g.*, sadness, stress, loss of interest and concentration.

Fuzzy Cognitive Maps (FCMs) consist a soft computing technique that has been used in many applications of several domains, including medicine (3). In general, the manual development of an FCM requires the participation of at least one specialist with experience. However, in some cases, no specialist may be available to help define an FCM, whereas it is difficult to manually design the model. In this paper, a novel framework for automatically creating FCMs is proposed, and is applied for depression severity estimation among adolescents. The introduced Constructive Fuzzy Cognitive Map (CFCM) utilizes features extracted using the recently proposed Constructive Fuzzy Representation Model (CFRM) (4). Specifically, CFRM detects the cause-and-effect relationships between the brain activity and the different states of depression, in an intuitive way, whereas it contributes to select the most informative features; thus, reducing the overall dimensionality of the problem under investigation.

[1] Corresponding Author, E-mail: diakovidis@uth.gr.

2. Material and Methods

To examine the brain, a segmentation of the EEG electrode positions into 11 regions was performed, based on (5). The electrodes were grouped into left, right, and midline frontal (LF, RF, MF), central (LC, RC, MC), and parietal (LP, RP, MP) regions (Figure 1).



1. Left Frontal (LF) {AF_3,F_1, F_3, F_5, F_7, FC_1, FC_3, FC_5 ,F_{p1}}
2. Left Central (LC) { C_1, C_3, C_5, CP_1, CP_3, CP_5 }
3. Left Parietal (LP) {P_1, P_3, P_5, P_7, PO_3, PO_5 ,PO_7 ,O_1}
4. Left Lateral (LL) { FT_7, T_7, TP_7 }
5. Midline Frontal (MF) { F_z, FC_z ,F_{pz} }
6. Midline Central (MC) { CP_z, C_z }
7. Midline Parietal (MP) { P_z, PO_z ,O_z }
8. Right Frontal (RF) { AF_4,F_2, F_4 , F_6, F_8, FC_2, FC_4, FC_6 , F_{p2}}
9. Right Central (RC) { C_2, C_4, C_6, CP_2, CP_4, CP_6 }
10. Right Parietal (RP) { P_2, P_4, P_6, P_8, PO_4, PO_6 , PO_8, O_2 }
11. Right Lateral (RL) { FT_8, T_8, TP_8 }

Figure 1. Segmentation of the electrical activity of the brain into 11 regions.

2.1 Constructive Fuzzy Representation of EEG Signals

Let us consider a feature vector $F_{r,\lambda}^{u} = \left(f_1^u, f_2^u, \dots, f_{N_r}^u\right)$, where u represents the wave types and $\lambda = 1,2, \dots, \Lambda$ is a class identifier corresponding to the degree of depression, *e.g.,* minimal depression; $r = 1,2, \dots, R$ are the examined brain regions, whereas N_r represents the number of EEG electrodes in the region r (Figure 1). Specifically, based on (5), a total number of $R = 11$ brain regions were used for the experiments. CFRM (4) is utilized to construct fuzzy sets and select the most informative features, which in our case correspond to the examined electrodes. The application of CFRM proceeds in 4 steps (Figure 2): a) it applies a clustering algorithm to group $F_{r,\lambda}^{u}$ into a set of M clusters with $M < K_\lambda$, where K_λ is the number of patients with $\lambda = 1,2, \dots, \Lambda$; b) The resulting centroids $q_m, m = 1, \dots, M$ and standard deviations $s_m, m = 1, \dots, M$ of the clusters are used to define the fuzzy sets; c) These fuzzy sets are then aggregated using the fuzzy union operation, resulting into new fuzzy sets, each of which corresponds to a feature f_n^u, n=1,2,…, N_r, and it has a membership function $\mu_\lambda(f_{k,n}^u)$, k =1,2,…, K_λ; d) All the membership functions $\mu_\lambda(f_{k,n}^u)$, k=1,2,…, K_λ , n=1,2,…, N_r, for $\lambda = 1,2, \dots, \Lambda$, are evaluated with respect to their overlap, and the features that correspond to fuzzy sets with highly overlapping membership functions are considered redundant and they are discarded. The application of this methodology on EEG features (6), showed that the delta and beta waves of all the examined electrodes, as well as those belonging to the lateral regions, *i.e.*, LL, RL, are redundant and they were discarded. Therefore, a total of 9 out of the 11 brain regions are considered for the experiments of this paper. In the sequel, for a given patient K the selected features $f_{K,n}^u, n = 1,2, \dots, N_r'$ ($N_r' \leq N_r$), $r = 1,2, \dots, 9$ are provided as input to the membership functions of the derived fuzzy sets, *i.e.*, the values $\mu_\lambda\left(f_{K,n}^u\right)$, $n = 1,2, \dots, N_r'$, $\lambda = 1,2, \dots, \Lambda$, are calculated. Considering

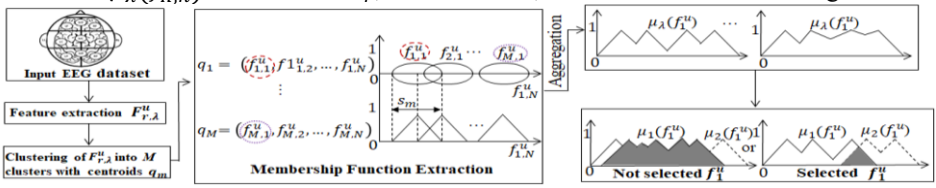


Figure 2. Overview of the Constructive Fuzzy Representation Model

$M_{K,r} = \max\left(\left(\mu_1\left(f^u_{K,n}\right), \mu_2\left(f^u_{K,n}\right), \mu_3\left(f^u_{K,n}\right)\right)\right)$ and $m_{K,r} = \left(\mu_1\left(f^u_{K,n}\right), \mu_2\left(f^u_{K,n}\right), \mu_3\left(f^u_{K,n}\right)\right)$, where $M_{K,r}$ corresponds to the maximum membership value of the selected fuzzy set $m_{K,r}$, the initial values A_r of the nodes of the CFCM are calculated by Eq.(1), which facilitates normalization purposes:

$$A_{K,r} = \begin{cases} 0 \le M_{K,r} \cdot \frac{1}{3} \le 0.33, & if\ m_{K,r} = 1 \\ 0.33 \le M_{K,r} \cdot \frac{2}{3} \le 0.66, & if\ m_{K,r} = 2 \\ 0.66 \le M_{K,r} \le 1, & if\ m_{K,r} = 3 \end{cases} \quad (1)$$

2.2 Constructive FCM

FCMs represent knowledge through concepts and directed, weighted edges between them (7). An FCM is defined as an ordered pair $< C, W >$, where C is the set of concepts and W is a quadratic matrix consisting of w_{ij} weights that determine the relationships among the concepts. The concept values of nodes $C = C_1, C_2, \ldots, C_n$, where n is the number of concepts, represent the state vector $A = \{A_r\}$.

The proposed CFCM is an FCM that is automatically constructed, given a set of initial concepts and a respective training dataset. In this study, a total of 11 initial concepts, representing the brain regions illustrated in Figure 1, were considered. The CRFM methodology is applied, and as described in section 2.2, from the 11 initial brain regions, only 9 of them are selected, and used for the construction of the CFCM model. Figure 3(a) illustrates the selected concepts of CFCM; C_1=Depression (D), which is the output concept, and the input concepts C_2= Left Central (LC), C_3= Left Frontal (LF), C_4= Left Parietal (LP), C_5= Right Central (RC), C_6= Right Frontal (RF), C_7= Right Parietal (RP), C_8= Midline Central (MC), C_9= Midline Frontal (MF) and C_{10}= Midline Parietal (MP). To estimate the weight matrix of CFCM (Figure 3(b)) the causal relationships between the concepts are examined. Initially, the average value $\overline{A_{r|\lambda}}$, $r = 1,2, \ldots, 9$, is obtained from all patients belonging to class λ in the training set, *i.e.*, the average of $A_{k,r}$, k=1,2,…, K_λ, for $\lambda = 1,2, \ldots, \Lambda$. This average value characterizes the brain activity of the respective depression state c. Then, the influence between two concepts C_i and C_j, $i \neq j$, which represent the brain activity in two brain regions i and j, can be defined with respect to the differences observed in the brain activity in all cases of depression, as $E^{\lambda_2,\lambda_1}_{i\to j} = (\overline{A_{\iota|\lambda_2}} + \overline{A_{j|\lambda_2}} - \overline{A_{\iota|\lambda_1}} - \overline{A_{j|\lambda_1}})/(\overline{A_{\iota|\lambda_2}} + \overline{A_{j|\lambda_2}} + \overline{A_{\iota|\lambda_1}} + \overline{A_{j|\lambda_1}})$, where $i, j = 1,2, \ldots, 9$, and $\lambda_1, \lambda_2 = 1,2, \ldots, \Lambda, \lambda_1 < \lambda_2$. The computed $E^{\lambda_2,\lambda_1}_{i\to j}$ are subsequently fuzzified using fuzzy sets defined in $[-2 \cdot \min(E_{i\to j}), 2 \cdot \max(E_{i\to j})]$ (Figure 3(c)). The fuzzy set in which $E^{\lambda_2,\lambda_1}_{i\to j}$ exhibit the maximum membership, is selected to linguistically represent the respective influence, using one of the three linguistic values, "negative", "neutral" and "positive". The final weight of each edge $i \to j$ is calculated as the center of gravity of the membership function obtained by the aggregation (using the algebraic sum) of the respective membership functions of $E^{\lambda_2,\lambda_1}_{i\to j}$. For a test case of a patient with unknown severity of depression, the developed CFCM iteratively calculates its state until convergence, for T iterations, according to the equation $A_i^{t+1} = g\left(A_i^t + \sum_{j=1, j\neq i}^{n} w_{ji} A_j^t\right)$, where $t = 1 \ldots, T$ is the iteration, w_{ji} is the weight matrix of the edge connecting C_j to C_i, and g is the log sigmoid function. The values of the initial state vector $A^0 = (A_1^0, \ldots, A_n^0)$ are estimated by Eq.(1) using EEG measurements from the test patient.

2.3 Data Description and Preprocessing

The dataset used in the experiments provides behavioral and electrophysiological information during eyes-closed and -open sessions over 60 electrode sites from 85 adolescents with minimal, mild, and moderate depression (6). In addition, the dataset was smoothed and averaged over five wave types: delta (1–4 Hz), theta (4–8 Hz), lower alpha (8–10 Hz), upper alpha (10–12 Hz), and beta (13–30 Hz). The absolute power of the EEG for five frequency bands was considered, resulting in a total of 5×60=300 EEG features. A total of 10 variations of this dataset was created, by following a 10-fold cross validation of the patient samples into non-overlapping training and test subsets.

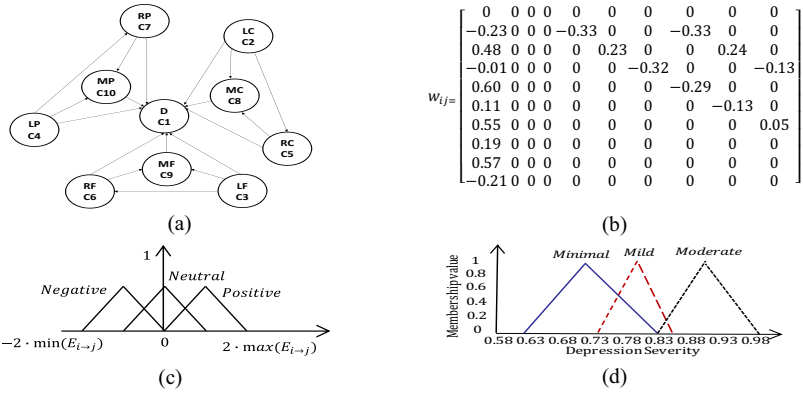
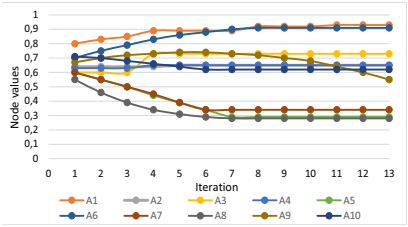


Figure 3. (a) CFCM structure. (b) Weight matrix of CFCM. (c) Membership functions of the influences between the CFCM concepts. (d) Membership functions for depression severity (output concept C_1).

3. Depression Severity Evaluation

As an example of the test phase described in the previous section, patient 76 of the available dataset was randomly selected. Using Eq.(1) the initial state vector $A^0 = (0, 0.66, 1, 0.66, 0.66, 0.64, 0.66, 0.64, 0.58, 0.97)$ is calculated. CFCM converged to a steady state after 13 iterations, using the weight matrix of Figure 3(b). The resulting state vector was $A^{13} = (\mathbf{0.93}, 0.73, 0.65, 0.09, 0.97, 0.15, 0.58, 0.55, 0.36)$ which shows that the patient has a "Moderate" depression equal to 0.93, taking into consideration Figure 3(d). According to A^{13}, a depressive patient has a more intense electrical activity on the left than on the right hemisphere of the brain. This can be inferred from the fact that $LC = 0.65 < RC = 0.09$, $LF = 0.73 < 0.97$ and $LP = 0.65 < RP = 0.10$. The evolution of the concept values, during the iterations, is illustrated in Figure 4(a).

Considering that this dataset has not been previously used in a classification context, we evaluated its decision making performance in comparison to four well-known classifiers (4). As it can be observed from Figure 4(b), CFCM provides higher or comparable results with significantly lower dimensionality (Dim), to the compared classifiers. Also, CFCM provides outcomes that are easily interpretable, based on its graph, explaining causal relationships of the involved concepts.



(a)

Classifier	Accuracy	Precision	Recall	Dim
SVM	0.64	0.69	0.62	300
k-NN	0.56	0.59	0.55	300
Naïve Bayes	0.58	0.65	0.58	300
CFRM	0.64	0.69	0.62	62
CFCM	**0.65**	**0.75**	**0.65**	**9**

(b)

Figure 4. (a) Values of CFCM nodes for a depressive patient. (b) Comparisons of CFCM with other classifiers.

4. Conclusion

In this paper, a novel framework for automatically constructing an FCM was proposed, with application to the depression severity among adolescents, based on their EEG. The experiments demonstrated the effectiveness of the introduced CFCM model, as it succeeded in identifying which brain regions were most associated with depression, while automatically detecting the interconnections between them. CFCM is capable of providing easily interpretable results, while being aware of uncertainty, and it is simple to implement. Future work includes further investigation of the proposed framework, using different types of membership functions and its application on different contexts.

Acknowledgment

We acknowledge support of this work by the project “Smart Tourist” (MIS 5047243) which is implemented under the Action “Reinforcement of the Research and Innovation Infrastructure”, funded by the Operational Program "Competitiveness, Entrepreneurship and Innovation" (NSRF 2014-2020) and co-financed by Greece and the European Union (European Regional Development Fund).

References

1. World Health Organization (WHO), others. Depression and other common mental disorders: global health estimates. Document, World Health Organization; 2017.
2. Field T, Miguel D, Sanders C. Adolescent depression and risk factors. Adolescence. Libra Publishers Incorporated; 2001;36(143):491.
3. Amirkhani A, Papageorgiou EI, Mohseni A, Mosavi MR. A review of fuzzy cognitive maps in medicine: Taxonomy, methods, and applications. Computer methods and programs in biomedicine. Elsevier; 2017;142:129–145.
4. Vasilakakis MD, Iakovidis DK, Koulaouzidis G. A Constructive Fuzzy Representation Model for Heart Data Classification. Public Health and Informatics. IOS Press; 2021. p. 13–17.
5. Kielar A, Joanisse MF. Graded effects of regularity in language revealed by N400 indices of morphological priming. Journal of cognitive neuroscience. MIT Press One Rogers Street, Cambridge, MA 02142-1209, USA journals-info …; 2010;22(7):1373–1398.
6. Rachamanee S, Wongupparaj P. Resting-state EEG datasets of adolescents with mild, minimal, and moderate depression. BMC Research Notes. Springer; 2021;14(1):1–3.
7. Kosko B. Fuzzy cognitive maps. International journal of man-machine studies. Els.; 1986;24(1):65–75.

Challenges of Trustable AI and Added-Value on Health
B. Séroussi et al. (Eds.)

doi:10.3233/SHTI220507

A Federated and Distributed Data Management Infrastructure to Enable Public Health Surveillance from Intensive Care Unit Data

Jonas BIENZEISLER[a,1], Lucas TRIEFENBACH[a], Alexander KOMBEIZ[a], Matthäus LOTTES[b], Christopher VOGEL[b], Linus GRABENHENRICH[b], Martina FISCHER[b], Theresa KOCHER[b], Lukas NIEKRENZ[c], Michael DREHER[c], Christoph MÜLLER[d], Rainer RÖHRIG[a] and Raphael W. MAJEED[a] on behalf of the AKTIN and SPoCK Research Group

[a]*Institute of Medical Informatics, Medical Faculty of RWTH Aachen, Aachen, Germany*
[b]*Department of Methodology and Research Infrastructure, Robert Koch Institute, Berlin, Germany*
[c]*Department of Pulmonology and Intensive Care Medicine, Medical Faculty of RWTH Aachen, Aachen, Germany*
[d]*Data Integration Centre (DIC), University Hospital RWTH Aachen, Aachen, Germany*

Abstract. The Robert Koch Institute (RKI) monitors the actual number of COVID-19 patients requiring intensive care from aggregated data reported by hospitals in Germany. So far, there is no infrastructure to make use of individual patient-level data from intensive care units for public health surveillance. Adopting concepts and components of the already established AKTIN Emergency Department Data registry, we implemented the prototype of a federated and distributed research infrastructure giving the RKI access to patient-level intensive care data.

Keywords. Critical care, routine care research, EHR, database, data privacy

1. Introduction

Intensive Care Units (ICU) are not only a crucial part of the health system but also an essential data source for public health surveillance. For the modeling and prediction of ICU occupancy during the COVID-19 pandemic, German authorities rely on mandatory reporting of COVID-19 cases and aggregated capacity data from hospitals [1]. Such data are collected and analyzed by the Robert-Koch Institute (RKI), the German National Public Health Institute. In the Project SPoCK (*Steuerungs-Prognose von intensivmedizinischen COVID-19-Kapazitäten*), the RKI, cooperating with universities, aims at improving the prediction of intensive care COVID-19 capacities. The automated provision of routine data sources on an individual-patient level is expected to enhance the predictive precision of the models. Furthermore, such data is seen of great value to

[1] Jonas Bienzeisler, Corresponding author; E-mail: jbienzeisler@ukaachen.de

improve *public health surveillance* in general [2]. Internationally, routine data could also be re-used for comparative effectiveness research or evaluation of COVID-19 government interventions. However, assessment of data on individual-patient level from multiple sites is opposed by legal guidelines in the European Union. The General Data Protection Regulation (GDPR) limits the sharing of sensitive data, typically requiring explicit patient consent for data processing. Our objective was to develop and operate a GDPR-compliant research infrastructure for routine data stemming from ICUs, that enables public health surveillance and other applications through the RKI.

1.1. Background

In Germany, data from the intensive care registry[2] is used to track the number of COVID-19 patients requiring intensive care [1]. Aside from structural characteristics of hospitals that maintain intensive care beds, information on ICU capacities and COVID-19 case numbers are collected in a *central* database. For reimbursement, it is also mandatory for ICUs to collect vital parameters and multiple clinical scores in patient data management systems. In practice, clinical scores are used to assess the current, disease-related patient status for individual risk prediction and necessary resources [3]. The predictive value of these scores has also been demonstrated in patients with COVID-19 admitted to the ICU [4] and could be used in occupancy models for the management of ICU capacities [2].

These data cannot be used so far because centralized storage of individual health data under the GDPR typically requires consent which is not feasible for ICU patients and might cause a selection bias. Privacy-preserving methods are usually used to access and analyze sensitive medical data for research despite data protection regulation. Two strategies are common to make use of data; either algorithms are distributed to the decentralized data and only anonymous results are aggregated [5,6] or only data that can be considered anonymous is aggregated (as in the intensive care registry) [7,8].

The *AKTIN* (*Alliance for information and communication technology in intensive* care *and emergency medicine*) *Emergency Department Data Registry* demonstrates, that such an evaluation is possible using a federated and distributed research infrastructure based on *decentral* data warehouses for anonymous aggregation or distributed computing [7–9]. Employing the concepts and software solutions of the AKTIN Registry, we implemented a data management infrastructure for the evaluation of ICU routine records for health surveillance, the so-called *SPoCK Data-Infrastructure*.

2. Methods

The methods for concepting the SPoCK Data-Infrastructure are based on the specific requirements for data capture and general requirements for sustainable operation.

2.1. Specific Requirements for data capture

The RKI required fast and actual access to patient-level data from multiple German ICUs for public health surveillance of COVID-19. The central analysis of structured reports

[2] www.intensivregister.de

by the RKI was mandatory, as the data needs to be analyzed together with pre-existing surveillance models. Thus, standardized processes for retrieval, transmission, and analysis of collected data were necessary. Semantically and syntactically comparable and compatible data were required. Finally, participating ICUs needed the tooling and implementation guidelines to store data and render it accessible for central pooling.

2.2. General Requirements

The handling of the health data is subject to legal and ethical constraints varying from country to country. In the European Union and under the GDPR, consent is mandatory for processing health data, if opening clauses are not in place. Under the premise of proper technical and organizational measures, health data may be processed for reasons of public interest such as health surveillance (Art. 9 (2) (i), GDPR). Integration into national and international research efforts is mandated for sustainable infrastructure projects. The generic concepts of the telematics platform for medical research networks [10] and the German medical informatics initiative [11] determine a national state of the art. Common technologies, terminologies, and data models are obligatory for integration into international research efforts.

3. Results

The general and specific requirements of the data collection match the requirements of the AKTIN Registry [9]. For setting up the SPoCK Data-Infrastructure in a short time, components of the AKTIN Registry were therefore modified to suit the use case. As there were only minor resources for data collection in participating hospitals, we restricted data to routine records collected for billing purposes. We aligned our concepts with recommendations for data protection for medical research networks in Germany [10].

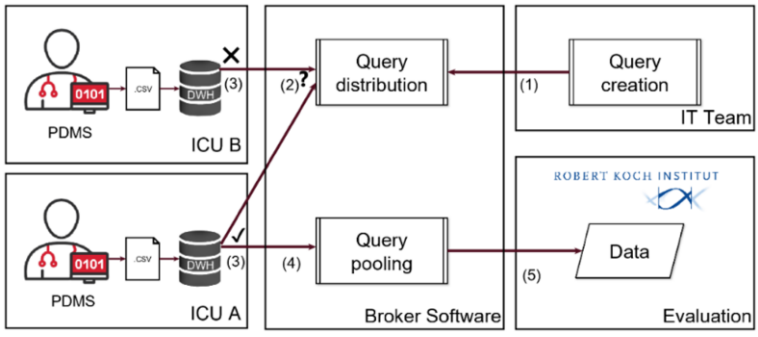


Figure 1: Federated and distributed research infrastructure of the Project. (1) The IT Team creates a query. (2) The query is distributed by the broker software to participating ICUs and can be checked (3) locally. (4) The SPoCK data warehouse (DWH) submits the results of the query to the broker software, anonymized data exports are pooled. (5) Data are delivered to the RKI.

3.1. Concept

We used the pre-existing software solutions of the AKTIN infrastructure for distributed data collection and federated data storage (c.f. Fig. 1). The used components are *not* connected to the AKTIN Registry but form a separate infrastructure. Data are stored in

modified instances of AKTIN data warehouses that employ an individual *SPoCK import script* [7,9]. The import script consists of an *Extract-Transform-Load* (ETL) pipeline employing Logical Observation Identifiers Names (LOINC) and Codes and Systematized Nomenclature of Medicine Clinical Terms (SNOMED CT) for annotation in i2b2. A modified instance of the AKTIN Broker can be used to query data. The data warehouse software is supplied and deployed using Linux-based installation packages.

Using these software components, a data set can be collected continuously and prospectively for the purposes of public health surveillance in a data warehouse. For simplicity, required data are exported in a CSV format from the patient data management system daily and then imported into the data warehouse. Each clinic operates and administrates the data warehouse on a dedicated server. The ownership of data and responsibility lies with the respective clinic. For public health surveillance, the data can be queried centrally via the broker and then evaluated by the RKI. It is the obligation of participating clinics to ensure that queried data are anonymous (c.f. Fig. 2).

3.2. Implementation

Currently, the framework is being piloted in the first clinic after approval of the ethics committee[3]. Data from 13144 cases were imported into the data warehouse. The first delivery of data to evaluation is intended for the first quarter of 2022. Furthermore, it is planned to roll out the software to two more clinics operating patient data management systems from different vendors.

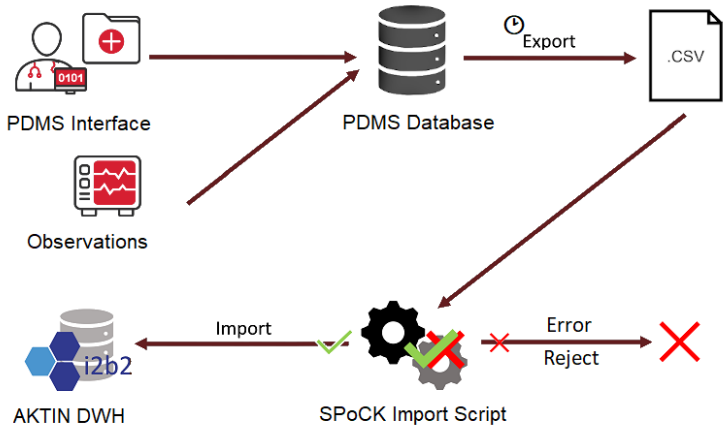


Figure 2: Data collection. Data are exported from patient data management system in CSV format using a repeating Cronjob. Exports are automatically imported into a AKTIN data warehouse (DWH) using the SPoCK import script which consists of an Extract-Transform-Load (ETL) pipeline.

4. Discussion and conclusions

The primary objective of our work was to enable public health surveillance from patient-level ICU data. We designed a research infrastructure that is capable of automatically providing the RKI with such data. We started with data collection and proved feasibility

[3] Ethics Committee Medical Faculty of RWTH Aachen University, EK 483/21.

in practice. The infrastructure is in operation. We based the infrastructure on pre-existing software solutions of the AKTIN Registry. The proof was given that the concepts and software components of the AKTIN Registry can be adapted to other scenarios and are ready for use in scenarios besides the emergency department. Further, the AKTIN components carry the potential for the rapid set up of health surveillance protocols. One advantage is that neither patients' informed consents need to be collected nor data needs to be aggregated on client side. As a result, the infrastructure can be scaled more easily to include data from small and large hospitals nationwide. The only prerequisite is automatic data provision and the installation of an AKTIN data warehouse. As long as the same terminologies and data models are used. it is possible to transfer the approach to other countries; import scripts or interfaces must be adapted accordingly.

Due to financial, organizational, and technical limitations, our work is limited to data collected from one ICU. Multiple instances have yet to be connected to a research network to provide data from multiple sites. Currently, data only consists of clinical scores collected for billing purposes. In the future, the data set needs to be extended and standardized to guarantee syntactic and semantic interoperability. Industry standards like HL7 ORO messages or HL7 FHIR resources should be used instead of CSV format to allow data stemming from any data source, nationally and internationally. Data are only collected for public health surveillance and cannot be used for health research. However, the infrastructure could be adapted to allow for potential health research as well.

References

[1] Schuppert A, Theisen S, Fränkel P, Weber-Carstens S, Karagiannidis C. Bundesweites Belastungsmodell für Intensivstationen durch COVID-19. Med Klin Intensivmed Notfmed 2021.

[2] Azzopardi-Muscat N, Kluge HHP, Asma S, Novillo-Ortiz D. A call to strengthen data in response to COVID-19 and beyond. J Am Med Inform Assoc 2021;28:638–39.

[3] Rapsang AG, Shyam DC. Scoring systems in the intensive care unit: A compendium. Indian J Crit Care 2014;18:220–28.

[4] Taylor EH, Marson EJ, Elhadi M, Macleod KDM, Yu YC, Davids R, et al. Factors associated with mortality in patients with COVID-19 admitted to intensive care: a systematic review and meta-analysis. Anaesthesia 2021;76:1224–32.

[5] Beyan O, Choudhury A, van Soest J, Kohlbacher O, Zimmermann L, Stenzhorn H, et al. Distributed Analytics on Sensitive Medical Data: The Personal Health Train. Data Intellegence 2020;2:96–107.

[6] Kapsner LA, Kampf MO, Seuchter SA, Gruendner J, Gulden C, Mate S, et al. Reduced Rate of Inpatient Hospital Admissions in 18 German University Hospitals During the COVID-19 Lockdown. Front Public Health 2020;8:594117.

[7] Brammen D, Greiner F, Kulla M, Otto R, Schirrmeister W, Thun S, et al. The German Emergency Department Data Registry – real-time data from emergency medicine: Implementation and first results from 15 emergency departments with focus on Federal Joint Committee's guidelines on acuity assessment. Med Klin Intensivmed Notfmed 2020.

[8] Boender TS, Cai W, Schranz M, Kocher T, Wagner B, Ullrich A, et al. Using routine emergency department data for syndromic surveillance of acute respiratory illness before and during the COVID-19 pandemic in Germany, week 10-2017 and 10-2021, Preprint at medrxiv 2021.

[9] Ahlbrandt J, Brammen D, Majeed RW, Lefering R, Semler SC, Thun S, et al. Balancing the need for big data and patient data privacy--an IT infrastructure for a decentralized emergency care research database. Stud Health Technol Inform 2014;205:750–54.

[10] Pommerening K., Helbing K, Ganslandt T, Drepper J Leitfaden zum Datenschutz in medizinischen Forschungsprojekten: Generische Lösungen der TMF 2.0. Schriftenreihe der TMF – Technologie- und Methodenplattform für die vernetzte medizinische Forschung e.V, Berlin: MWV Medizinisch Wissenschaftliche Verlagsgesellschaft mbH & Co. KG; 2017.

[11] Semler SC, Wissing F, Heyder R. German Medical Informatics Initiative: A National Approach to Integrating Health Data from Patient Care and Medical Research. Methods Inf Med 2018;57:e50-6.

Challenges of Trustable AI and Added-Value on Health
B. Séroussi et al. (Eds.)

doi:10.3233/SHTI220508

Minor and Parental Access to Electronic Health Records: Differences Across Four Countries

Josefin HAGSTRÖM [a,1], Isabella SCANDURRA [b], Jonas MOLL [b], Charlotte BLEASE[c], Barbara HAAGE [d], Iiris HÖRHAMMER [e], and Maria HÄGGLUND [a]
[a] *Healthcare Sciences and E-health, Dept Women's and Children's Health, Uppsala University, Sweden*
[b] *Informatics, School of Business Örebro University, Örebro, Sweden*
[c] *General Medicine and Primary Care, Beth Israel Deaconess Medical Center, Harvard Medical School, Boston, MA, USA*
[d] *Dept of Health Technologies, Tallinn University of Technology, Estland*
[e] *Dept of Industrial Engineering and Management, Aalto University, Espoo, Finland*

Abstract. An increasing number of countries are implementing patient access to electronic health records (EHR). However, EHR access for parents, children and adolescents presents ethical challenges of data integrity, and regulations vary across providers, regions, and countries. In the present study, we compare EHR access policy for parents, children and adolescents in four countries. Documentation from three areas: upper age limit of minors for which parents have access; age at which minors obtain access; and possibilities of access restriction and extension was collected from Sweden, Norway, Finland, and Estonia. Results showed that while all systems provided parents with automatic proxy access, age limits for its expiry differed. Furthermore, a lower minimum age than 18 for adolescent access was present in two of four countries. Differences between countries and potential implications for adolescents are discussed. We conclude that experiences of various approaches should be explored to promote the development of EHR regulations for parents, children and adolescents that increases safety, quality, and equality of care.

Keywords. Adolescents; Children, Parents; Patient accessible electronic health records; International comparison

1. Introduction

Implementation of patient accessible electronic health records (PAEHR) is ongoing globally, with around 20 countries developing online systems [1]. PAEHR systems enable access for patients to health information such as notes, medication, lab results etc. However, there is considerable variety in appearance and functionality of services used to provide patients with continuous health data access. In response to this rapid development, legal frameworks are continuously being adapted to improve use and ensure privacy of such PAEHR systems.

[1] Josefin Hagström, Department of Women's and Children's Health, Uppsala University, Dag Hammarskjölds väg 14B, 751 85 Uppsala, Sweden; E-mail: josefin.hagstrom@kbh.uu.se.

A key issue for implementing PAEHR is the area of parents, children, and adolescents. Though parents report benefits from having access to their child's EHR [2], issues around information privacy may arise as the child becomes an adolescent. As such, the adoption of EHR access in a pediatric population has raised ethical and legal challenges and been subject to much dispute [3]–[5]. Among concerns expressed by healthcare professionals (HCPs) regarding PAEHRs, it is anticipated that adolescents may have poor understanding of their clinical information which can lead to harm. In contrast, surveyed adolescents express a desire for access to their EHR [6] and a cross-sectional survey has revealed high satisfaction and adequate comprehension among adolescents [7]. Still, the transition from parental access to self-access requires careful consideration, as adolescents' acceptability of parental EHR access may be contingent on the relationship with the parent [8]. Furthermore, two US studies found adolescents to be less likely to share information with HCPs if unsure about confidentiality [9], [10].

There is not yet consensus about when to retract parental access and provide individuals with self-access, leaving a variability in policies across countries [1]. Some have adopted a case-by-case approach, often relying on subjective judgments about adolescents' maturity. To increase equality and furthermore, allow for exceptions, a need for shared principles has been suggested, rather than determined standards [1]. Development of such principles should be informed by current PAEHR policies. Therefore, the aim of the study is to compare national PAEHR regulations and services for parental access and adolescent self-access to EHR in four countries. We focus on three areas: upper age limit of minors for which parents have access, age at which minors obtain access, and possibilities of access restriction and extension.

2. Methods

The study was carried out within the NORDeHEALTH research project [11]. In project meetings, workshops were planned for designing a socio-technical comparison of PAEHR implementation in Sweden, Norway, Finland, and Estonia. The selected countries are part of the project and target of this study due to being forerunners in PAEHR implementation. A socio-technical analysis was carried out, informed by a framework by Sittig and Singh [12] especially tailored to health systems. For each dimension proposed in the framework a number of questions were developed to cover socio-technical situations regarding development, implementation and use of patient portals. Analysis categories related to the eight framework dimensions *Hardware & software computing infrastructure; Clinical content; Human computer interface; People; Workflow & communication; Internal organizational features; External rules, regulations & pressures;* and *System measurement & monitoring*, and the added dimension *Features & functions* were continuously refined during several workshops between 3/2021 to 5/2021 with between one to four topic experts from each of the countries mentioned above. This process resulted in a data collection form with questions to be answered by the respective topic experts in relation to their own country's PAEHR system(s). This paper focuses on a subset of questions from the dimension *External rules, regulations & pressures,* that were of special interest regarding parental and adolescent access to records, namely: 'At which age do minors obtain access to their own PAEHR?'; 'Do parents have the right to access their children's PAEHR?'; 'Is any potentially sensitive health information exempted from parental access?'; 'Can minors apply for restrictions in parental access?'; 'Can parents apply for prolonged access to the child's

EHR beyond the default?'; and 'Can minors apply for earlier access to their EHR before the default set-up?'.

The original data collection was finished in November 2021 and a complementary, smaller, data collection was carried out during a workshop in December 2021 to elaborate on questions related to parental and adolescent access to records online in particular. The same topic experts in each country were surveyed and were contacted via e-mail when additional clarification was required. Analysis was performed by the primary author by comparing the answers to these questions across the four countries.

3. Results

All national systems in the four countries provided parents with automatic access to their child's EHR at birth. Figure 1 visualizes the patterns for parental EHR access and child/adolescent self-access in the four studied countries.

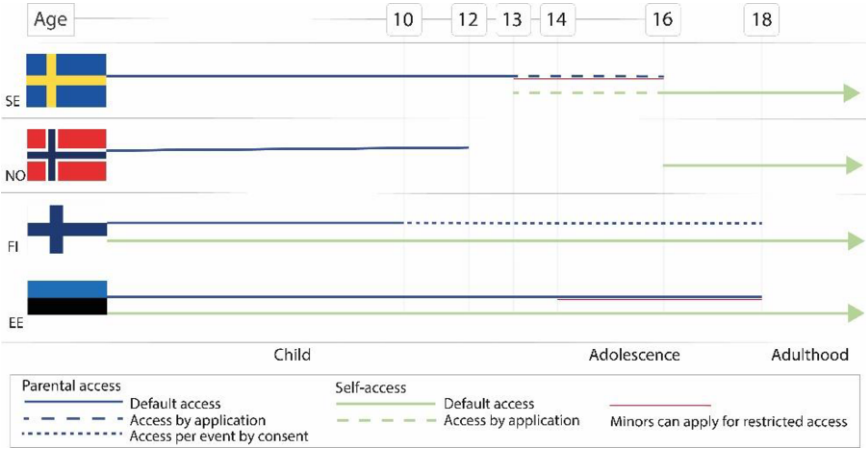


Figure 1. Parental and child/adolescent access to EHRs in Sweden, Norway, Finland and Estonia.

Age limits for parental loss of access were lower in Sweden (age 13) and Norway (age 12), than in Finland and Estonia (age 18). In Finland, parental access for minors above the age of 10 was not the default, though access was subject to the minor's capability to decide for their own care and their consent in the following way: for each care event and prescription, an HCP had to select one of four options: (1) the minor does not have decision-making capacity, and the event note is made accessible for parents; (2) the minor has decision-making capacity and consents to making the event note accessible for parents; (3) the minor has decision-making capacity and does not consent to making the event note accessible for parents; (4) the minor's decision-making capacity is unknown, and the event note is not made accessible for parents.

The age of obtaining EHR self-access was the same in Sweden and Norway (age 16) while by contrast in Finland and Estonia, there was no lower age for when the minor obtained access to their EHR. Instead, children and adolescents could access their own records when they had acquired an electronic ID. Sweden and Norway also required an electronic ID to access the national PAEHR, for which one must be 8 (Sweden) and 16 years old (Norway). None of the studied countries exempted any potentially sensitive health information by default from parent access. In all countries except for Norway,

adolescents were able to apply for restrictions in parental access to their EHR. In Estonia, the lower age limit for this was 14.

In Sweden and Norway, where parents' access to their child's record is by default ended as the child turns 16, there is a possibility for parents to apply for a prolonged access. Of the two, only parents in Sweden had the opportunity to apply for access after the child had turned 13 years old. Because the procedure relying on adolescents' capability to decide for their own care and their consent applies until the age 18 in Finland, there is no separate procedure for prolonging parental access. Similarly, the ability to apply for earlier access was only applicable for Sweden and Norway where there were lower age limits. Of these, only adolescents in Sweden could apply for self-access to their EHR prior to the age of 16, at which point they gained automatic access.

4. Discussion

Our findings show that the national eHealth services in Sweden and Norway held similarities in regards to regulations on parental access and self-access for minors, whereas Finland and Estonia had a somewhat different approach. Sweden and Norway used a default blocked access approach as soon as the child reached the age of 16, while Finland and Estonia enabled parent access until the child turned 18.

Finland held the lowest age limit (age 10) for parental loss of default access. Parental access for minors older than 10 in Finland was decided on event-basis, which prevents some of the risks in the approach of default shared access of parents and adolescents used in Estonia. One may argue that not all children aged between 10-14 can fully understand the information in their EHR and independently make informed decisions around their health, however, it is worth emphasizing that at least some researchers argue that children are capable of full informed consent from the age of 12 [13]. The ability to protect the privacy of adolescents is important as situations that may threaten the minor's wellbeing are numerous and hard to foresee. For example, adolescents may not want to reveal information pertaining to sexual activity, disclosure of alcohol or drug abuse, or stigmatized illnesses such as anxiety or depression. While shared access by event increases privacy, issues may still arise in cases when part of a visit concerns confidential information [5]. Furthermore, the regulations in Finland largely depend on subjective judgment of the professional treating the adolescent.

The more restrictive regulations in Sweden and Norway lead to other challenges: namely, a "gap" between the ages of 13-16 (from age 12 in Norway) where neither parent nor child can access the records. The age 13-limit is in fact not legislated but was set in 2012, and established in the national regulatory framework of the PAEHR in collaboration with the Swedish Association of Local Authorities and Regions (SALAR) [14], and thereafter Norway made similar decisions. This lack of access has been subject to criticism in Sweden [15], where parents ask for access to "non-sensitive" information. Though it is possible for parents and adolescents in Sweden to apply for access outside of regulations, the procedure requires knowledge, time, and effort.

Lastly, while none of the countries studied here had by regulation exempted any sensitive information in the child's EHR from parental access, implementation may vary in regards to setting and region. For example, access to records from psychiatry have been topic of considerable debate for both adult and youth psychiatry [16]. As of 2021, one of Sweden's 21 regions has decided to exempt child and adolescent psychiatry notes from availability in the national EHR service.

It should be noted that implementation of these services is ongoing and while all functionalities mentioned in this paper are available in each country, they may not yet have been implemented across all regional settings.

5. Conclusions

The complexity of PAEHRs in the context of the pediatric population is evident from its policy diversity across countries. This may be the first attempt to compare detailed PAEHR regulations of parental access and adolescent self-access to EHR. Findings gained from the four countries can be of assistance to future work with a greater focus on improving our understanding of how these approaches are experienced by stakeholders, primarily parents and young patients. Shared insights from a diversity of approaches necessitate concurrent safety monitoring. This will be indispensable to improve safety and equality of care in this emergent field.

References

[1] Essén A, Scandurra I, Gerrits R, Humphrey G, Johansen MA, Kierkegaard P, et al. Patient access to electronic health records: differences across ten countries. Health Policy Technol. 2018;7(1):44–56.
[2] Britto MT, Hesse EA, Kamdar OJ, Munafo JK. Parents' Perceptions of a Patient Portal for Managing Their Child's Chronic Illness. J Pediatr. 2013 Jul 1;163(1):280-1.e2.
[3] Sharko M, Wilcox L, Hong MK, Ancker JS. Variability in adolescent portal privacy features: How the unique privacy needs of the adolescent patient create a complex decision-making process. J Am Med Inform Assoc. 2018 Aug;25(8):1008–17.
[4] Stablein T, Loud KJ, DiCapua C, Anthony DL. The Catch to Confidentiality: The Use of Electronic Health Records in Adolescent Health Care. J Adolesc Health. 2018 May;62(5):577–82.
[5] Bourgeois FC, DesRoches CM, Bell SK. Ethical Challenges Raised by OpenNotes for Pediatric and Adolescent Patients. Pediatrics. 2018 Jun;141(6):2021.
[6] Ransom NR. A teen's perspective: Adolescent access to their own electronic medical records. Int J Adolesc Med Health. 2016;28(1):123–4.
[7] Huang JS, Yueh R, Ma S, Cruz R, Bauman L, Choi LJ. Adolescents' and Young Adults' Satisfaction with and Understanding of Medical Notes from a Pediatric Gastroenterology Practice: A Cross-Sectional Cohort Study. J Pediatr. 2019 Dec;215:264–6.
[8] Carlson JL, Goldstein R, Buhr T, Buhr N. Teenager, parent, and clinician perspectives on the electronic health record. Pediatrics. 2020 Mar;145(3). Available from: https://doi.org/10.1542/peds.2019-0193
[9] Klein JD, McNulty M, Flatau CN. Adolescents' access to care: Teenagers' self-reported use of services and perceived access to confidential care. Arch Pediatr Adolesc Med. 1998 Jul;152(7):676–82.
[10] Campos-Castillo C, Anthony DL. The double-edged sword of electronic health records: implications for patient disclosure. J Am Med Inform Assoc JAMIA. 2015 Jul;22(e1):e130–40.
[11] Sittig DF, Singh H. A New Socio-technical Model for Studying Health Information Technology in Complex Adaptive Healthcare Systems. Qual Saf Health Care. 2010 Oct;19(Suppl 3):i68–74.
[12] Hägglund M. Nordic countries lead new initiative on patient access to EHRs [Internet]. TheBMJopinion. 2022 [cited 2022 Jan 16]. Available from: https://blogs.bmj.com/bmj/ 2021/05/18/maria-hagglund-nordic-countries-lead-new-initiative-on-patient-access-to-ehrs/
[13] Hein IM, De Vries MC, Troost PW, Meynen G, Van Goudoever JB, Lindauer RJL. Informed consent instead of assent is appropriate in children from the age of twelve: Policy implications of new findings on children's competence to consent to clinical research. BMC Med Ethics. 2015 Nov 9;16(1):76
[14] Scandurra I, Lyttkens L, Eklund B. Implications of Swedish National Regulatory Framework of the Patient Accessible Electronic Health Record. Stud Health Technol Inform. 2016;228:695–9.
[15] Lernfelt M. 13-åringar kan inte förväntas sköta sin vård själva [Internet]. Dagens Samhälle. 2021 [cited 2022 Jan 20]. Available from: https://www.kristianstadsbladet.se/ledare/13-aringar-kan-inte-forvantas-skota-sin-vard-sjalva-cdb7872f/
[16] Petersson L, Erlingsdóttir G. Open Notes in Swedish Psychiatric Care (Part 2): Survey Among Psychiatric Care Professionals. JMIR Ment Health. 2018;5(2):e1052.

Challenges of Trustable AI and Added-Value on Health
B. Séroussi et al. (Eds.)

doi:10.3233/SHTI220509

National Integration Components Challenge the Epic Implementation in Central Norway

Gunnar ELLINGSEN[a,1], Morten HERTZUM[b], Bente CHRISTENSEN[c], Rolf WYNN[a]
[a] *UiT – The Arctic University of Norway, Tromsø, Norway*
[b] *University of Copenhagen, Copenhagen, Denmark*
[c] *Nord University, Bodø, Norway*

Abstract. Electronic health record (EHR) suites cover a broad range of cross-sectoral use scenarios. Thereby, they streamline information flows but also require that healthcare professionals with diverse responsibilities must adapt to one and the same system. In the region of Central Norway, the EHR suite from Epic is being implemented at hospitals as well as in municipal healthcare. However, the 64 municipalities in the region are increasingly exploring the option of bypassing Epic by supplementing their existing systems with national integration components. These components provide integration and data exchange across systems for selected healthcare information. We discuss whether they are a viable alternative to Epic. The three components are the summary care record, the shared medication list, and the national welfare technology hub.

Keywords. electronic health record, EHR suite, Epic, national components

1. Introduction

The overall goal of large-scale electronic health record (EHR) suites is to cover a broad range of use scenarios for healthcare workers in hospitals, nursing homes, home-care services, and general practitioner (GP) clinics. The EHR suite can then be used as the principal system in entire healthcare regions to ensure an efficient information flow among the various practices. However, large EHR suites have been increasingly criticized for their "one-size-fits-all" features, and thus the significant consequences for the various organizations involved, all of which must adapt to the same system [1]. Its implementation usually involves replacing most of the existing ICT portfolio, requires meticulous planning, and consumes considerable financial and human resources. This is of particular concern for municipalities, which are responsible for providing first-line healthcare services to citizens in the local community, such as managing nursing homes and home-care services. At the lowest politically-elected level, municipalities have fewer resources for their health services than the state-owned tertiary institutions. This includes limited resources for engaging in large-scale EHR suite implementations. As a result, they must carefully consider their strategy for ICT investments. On this basis, we ask the research *question: What viable alternatives to EHR suites do municipalities have for ensuring an efficient information flow among the different healthcare domains?* Empirically, we focus on the Health Platform program in Central Norway,

[1] Gunnar Ellingsen, Corresponding author, UIT - The Arctic University of Norway, research group: SOF NORD, 9038 Tromsø, Norway; E-mail: gunnar.ellingsen@uit.no.

where the goal is to implement the U.S. Epic EHR suite in 2022 [1,2]. While many municipalities in Central Norway have decided to participate, others are exploring the option of continuing to use their old systems by supplementing them with national integration components, which serve some of the same purposes as the Epic EHR suite. Theoretically, this paper is positioned in the information infrastructure literature [3]. We draw particularly on its installed base concept, which emphasizes that new functionality evolves on top of existing functionality rather than replaces it [4].

2. Method

Our study has an interpretive research approach, which considers a phenomenon from different viewpoints [5,6]. We compare the aim of the Health Platform program with the opinions of three municipality consortia that are considering alternatives to the Health Platform program. During spring of 2021, we conducted three one-hour interviews with four informants from three different municipality consortia in Central Norway. The informants have many years of professional experience with EHRs and with advising the top municipal leadership. We also include six interviews conducted with Health Platform managers in 2018 and nine interviews with GPs in Central Norway in 2019. The interviews were conducted in an open-ended manner, but in a broad sense, the focus in the interviews was on potentials and challenges with Epic.

3. Results

3.1. The Health Platform program in Central Norway

The Health Platform is a regional program jointly owned by the Central Norway Regional Health Authority and the Trondheim municipality. It is also a pilot for the national goal of "one citizen – one record". In 2019, the program signed a NOK 2.7 billion (EUR 270 million) contract with Epic Systems Corporation to acquire and implement the Epic EHR suite in Central Norway, including all hospitals, GP clinics, nursing homes, and home- care services. The number of healthcare workers in the region totals around 44,000. As a suite system, Epic is largely self-contained. Most of the functionalities needed by health personnel are supposed to be provided by Epic, either as ready-for-use functionality or through configuration. The Central Norway Regional Health Authority and the Trondheim municipality will start implementing Epic in the three regional hospitals and in the Trondheim municipality in 2022. Thereafter, the implementation process will continue in other municipalities that choose to join the program. While the hospitals in the region must replace their current EHR, the municipalities are in no immediate hurry. Although their current systems in nursing homes and home-care services lack the most recent functionality, there is no urgent need to replace them. Similarly, GPs have quite modern systems in place, with which they are satisfied. However, since the goal of the program is to create a comprehensive health service, the Health Platform strongly encourages all municipalities to participate to secure complete and up-to-date information about the condition, treatment and medication status of the patients. While almost half of the 64 municipalities in Central Norway have indicated that they will implement Epic, the rest have become increasingly reluctant. As go-live approaches, politicians in the municipalities are realizing that

implementing Epic will drain the municipalities' human and financial resources. Therefore, some municipalities are considering the so-called "zero-alternative," which implies continuing to use their current EHR systems and upgrading these systems with national integration components, which are in the pipeline.

3.2. The summary care record

The summary care record (in development since 2012) is a digital solution for sharing patients' health information across the health sector. By using it, healthcare professionals have access to the same information regardless of whether they work in a hospital, a GP clinic, or an emergency room. The summary care record is integrated with the EHRs in the sector and provides real-time access to patients' critical information, pharmacy-dispensed medication, discharge letters, and laboratory results (including COVID-19). In addition, the possibility of sharing various clinical documents from Norwegian hospitals is currently being tested at different locations in Norway. Citizens may register information in the record, such as their primary contact person, disease history (structured selections), and special needs. Healthcare personnel must actively choose what information from their local EHRs they will share in the summary care record. Our informants underscore the fact that while GPs have been slow to make information in their local EHR available to the summary care record – and thus have hindered its potential – many GPs have increasingly begun to share their information. The current COVID-19 pandemic has given the care record a boost due to how easily available laboratory results from COVID tests are to various health personnel. Thus, one of our informants was quite optimistic about its potential and made a favorable comparison with the Epic EHR: "When GPs start exchanging information through the summary care record where residents, specialist health services, and the emergency services all have access, I'm not sure that GPs need so much more [i.e., the Epic EHR]." Another informant from the municipality consortium said that they had recently implemented the summary care record in the consortium and were eagerly looking forward to utilizing its potential: "Obviously, you get a lot of information that is useful. Additionally, when you start sharing patient record documents in the summary care record you can have discharge letters, nursing documentation, physiotherapy notes, and whatever you like. You can also have laboratory results and X-ray results." Based on these prospects, our informants have started to reflect on what more they need, or if they really need anything more: "After we have selected everything that we want, we are left with the question of what we haven't got that is unique to Epic."

3.3. The shared medication list and the central prescribing module

The process of developing a shared medication list has been underway for a while, but from December 2021 it has been piloted in Norway's second-largest city, Bergen. Compared to the summary care record, which contains an overview of the pharmacy-dispensed medication, the shared medication list gives an overview of a patient's complete list of medications. This includes prescription medicines, non-prescription medicines, and medicines that have been administered in an emergency room, hospital, nursing home, or purchased abroad. In order to be able to introduce the shared medication list throughout the country, a national component called the Central Prescribing Module is also being developed and will be available from 2022. The Central Prescribing Module is a medication and requisition module that (through integration with the EHRs)

facilitates the sharing of medication information among various EHRs. Healthcare professionals who use the module will have a unified prescription user interface, regardless of which EHR they use. Eventually, the shared medication list will become part of the summary care record. All our informants expressed great faith in the shared medication list and questioned whether there were any benefits unique to Epic. One of them said, "When we did a survey in all the 64 opt-in municipalities, where we wanted to find out what benefits the municipalities were looking for with Epic, the shared medication list was number one. This was originally presented as something unique to Epic, so it was a fairly common misconception that Epic was the only way forward." Another informant said: "When we get the shared medication list integrated into the summary care record, then we will have a pretty good picture of the patient's health situation." A third informant said: "A shared medication list is not unique to Epic; it is a national functionality that all EHRs must comply with, including Epic. Epic must relate both to it and the central prescribing module." A positive factor is that the integration of EHRs with the national components can now be done more seamlessly than before, when the national components had to be integrated with all local installations. Now EHRs can be integrated with the central prescribing module located in the cloud, thereby ensuring that all municipalities are connected to the module at the same time.

3.4. The national welfare technology hub

For the most part, Epic is a hospital-oriented system, and therefore some functionality for municipalities must be developed. One key component is the integration with the national welfare technology hub, which enables integration between the municipal EHRs and their welfare technologies. This integration will ensure an efficient information flow among welfare technologies, EHRs, and response-center solutions. Unfortunately, Epic will not be able to deliver this integration in the first round of the project and, according to our informants, "perhaps not in the second round either." For municipalities, this is a serious concern. Many of them have been heavily engaged in the national welfare program that started in 2013, have invested a lot of resources and money in it, and "have come a long way in establishing working welfare solutions." These technologies include bed sensors, door sensors, digital monitoring, medication dispensers, and GPS tracking. A typical use case is when an alarm goes off at a patient's home, which then demands an action at the response center. Here, nurses or other care personnel monitor the situation for many patients through a response-center solution. Through integration with the municipal EHR, the health personnel have access to contact information, level of service provided, planned next visit, next of kin, etc. Our informants pointed out that this is crucial decision-support for the care staff in deciding how to respond. Losing the integration with the national welfare technology hub will mean that the municipalities suffer a serious setback in their welfare technology investments. Therefore, the cluster of municipalities requested that the Health Platform embed this integration in the plans for the implementation of Epic. However, this was not possible: "We started sending letters last summer to try to make them understand that this integration was important for the municipalities, but the planning for the implementation had already gone so far that if one integration was to go in, another had to come out and there was no room to take anything out." As a result, the municipalities are increasingly worried about not being listened to. As one of our informants put it: "It is clear that if there is something the hospital needs, then it is more likely to be perceived as more important than what the municipalities need."

4. Concluding discussion

Healthcare is delivered through a cross-sectoral collaboration that relies on a partially integrated technological infrastructure to solve problems central to citizens' wellbeing. There is a recognized need for more streamlined information flows and reduced data fragmentation. Large-scale EHR suites and national integration components are different approaches to meet this need. The former purports to replace large parts of the existing ICT portfolio with a single integrated system, and the latter to evolve the portfolio by integrating its systems better. While evolution is an incremental strategy that changes the technological infrastructure through a series of low-risk steps, replacement is a high-risk strategy because it severely upsets the equilibrium provided by the existing portfolio of systems. The information infrastructure literature emphasizes that this equilibrium consists of a wide-ranging network of interdependent relations that tend to remain unnoticed until they are disrupted [3,4]. That is, the replacement strategy is somewhat risky because its consequences are difficult to foresee. The three municipality consortia in our study are experiencing this difficulty. They lean toward a strategy based on integration components because it appears less drastic, less risky, and less costly. Their reservations about Epic highlight the challenges faced by the Health Platform program, which is responsible for the implementation of Epic and dependent on the participation of the municipalities to take full advantage of its functionality. It is increasingly evident that the national integration components will provide some of the same benefits as those initially ascribed exclusively to Epic. In the municipalities, this has prompted reflection about what the advantage of Epic really is. Such reflection does not just point to the possibility of some municipalities opting out of the Epic implementation, it undermines the rationale of implementing it in the first place, and thus the realization of the goal of the Health Program, that is, one EHR in Central Norway.

Acknowledgments

This study is part of the eMM and REAL projects, funded by the Research Council of Norway (grants no. 314382 and 326744).

References

[1] Hertzum, M, Ellingsen G. The implementation of an electronic health record: Comparing preparations for Epic in Norway with experiences from the UK and Denmark. International Journal of Medical Informatics. 2019; 129: 312–317.
[2] Ellingsen, G, Hertzum, M. User requirements meet large-scale EHR suites: Norwegian preparations for Epic. In: Pape-Haugaard, L, Lovis C, Madsen IC, Weber P, Nielsen PH, Scott P, editors. Proceedings of MIE; 2020 April 28 – May 1; Geneve: IOS Press.
[3] Aanestad M, Grisot M, Hanseth O, Vassilakopoulou P, editors. Information infrastructures within European health care: Working with the installed base. Cham, Switzerland:Springer; 2017. pp. 25-33.
[4] Star SL, Ruhleder K. Steps toward an ecology of infrastructure: Design and access for large information spaces. Information Systems Research. 1996; 7(1): 111-134.
[5] Klein H, Myers, M. A set of principles for conducting and evaluating interpretive field studies in information systems. MIS Quarterly. 1999; 2: 67-94.
[6] Walsham, G. Interpretive case studies in IS research. Nature and method. European Journal of Information Systems. 1995; 4: 74–81.

Challenges of Trustable AI and Added-Value on Health
B. Séroussi et al. (Eds.)

doi:10.3233/SHTI220510

Implementation of a Data Warehouse in Primary Care: First Analyses with Elderly Patients

Mathilde FRUCHART[a,1] Paul QUINDROIT[a], Haris PATEL[a], Jean-Baptiste BEUSCART[a], Matthieu CALAFIORE[a,b] and Antoine LAMER[a]
[a] *Univ. Lille, CHU Lille, ULR 2694 - METRICS: Évaluation des Technologies de santé et des Pratiques médicales, F-59000 Lille, France.*
[b] *Univ. Lille, Département de médecine générale, F-59000 Lille, France*

Abstract. The implementation of clinical data warehouses has advanced in recent years. The standardization of clinical data in these warehouses has made it possible to carry out multicenter studies and to formalize the clinical vocabulary. However, there is limited insight into a patient's overall care pathway in the clinical domain. Regarding primary care data, the implementation of this type of warehouse in a routine way is hindered in particular by the analysis of textual data provided by general practitioners during patient consultations. In our study we collected primary care data for standardization in a data warehouse. The purpose of this analysis was to assess the feasibility of analyzing primary care data, and particularly to study the consultations and prescriptions of the elderly patient contained in our primary care data warehouse.

Keywords. Primary care, Data warehouse, Elderly patient, Data Reuse, Electronic Health Record

1. Introduction

The digitalization of health facilities has enabled the automated collection of electronic health records primarily generated for care or administrative purposes [1]. This offers the possibility of a secondary use for research [2], particularly through the implementation of clinical data warehouses [3].

Data warehouses are mainly developed on the scope of hospitals or national claims. On one hand, in the hospital, they provide a complete view of patient management in the care units with the diagnoses, the procedures, the drugs administrations, the biology results and the discharge letters. However, they do not allow us to explore the patient outcomes after discharge and in other hospitals. On the other hand, national claims contain data collected anonymously and prospectively for all national health insurance beneficiaries [4]. They are dedicated to reimbursed inpatient and outpatient care (e.g., general practitioners [GPs], pharmacies, nurses, etc.) and do not include clinical data. New attempts have emerged for primary care [5,6]. They provide data the same patient

[1]Corresponding Author, Mathilde Fruchart, ULR 2694, Lille University, 2 place de Verdun, F-59000, Lille, France; E-mail: mathilde.fruchart@univ-lille.fr.

over the long term, such as weight, body mass index, results of blood tests and symptoms experienced by the patient during each medical visit [7].

In our study, we have implemented a data warehouse with data collected in a GPs center and we studied clinical and longitudinal data from the general population and the elderly.

2. Methods

We implemented a data warehouse with the GPs center of Wattrelos, in north of France. The center includes 6 general practitioners. The software Weda (WEDA, Montepellier, France) collects the data during the consultations, with the patient history and symptoms, the vital signs and biometric data (e.g., heart rate, arterial pressure, temperature, weight, height, body mass index), the drug prescriptions, the blood tests, and the billing data. Drug prescriptions are documented with the *Code Identifiant de Présentation* (CIP), a French terminology. CIP codes were mapped with Anatomical Therapeutic Chemical (ATC) codes. Patient history and symptoms are described with the International Classification of Diseases in the 10th version (ICD10) and free text. The blood tests results are documented after receiving the analysis reports from the laboratories, and the codes depend of the laboratory terminology. The vital signs and biometric data are specified with non-standardized questions. The mapping to the ATC, blood tests labels and questions has been conducted beforehand and verified by a team of pharmacists, geriatricians and GPs belonging to the research team. We removed all sensitive data (i.e., identity, birth date, patient's address) and we created a unique artificial identifier per patient. The data warehouse is stored on a secure server disconnected from the internet.

An extract-transform-load (ETL) process was implemented to automatically read XML files and integrated all the information listed in a relational database. The process will be relaunched each time we get a new extraction from the editor. In order to enhance reproducibility and to collaborate with other centers and other software, we implemented the Observational medical outcomes partnership Common data model (OMOP-CDM) [8]. The use of this CDM allows to reproduce and share data, methods and results across different centers. The ETL was implemented in Python and Postgresql. The use of OHDSI tools was applied in particular for vocabulary mapping (Atlas, Athena) [8].

We extracted all records of consultations that occurred between 2013 to 2020 with at least one drug prescription. The extraction results in a set of XML files for each patient with a hierarchy of information. An exploratory analysis was done to judge the quality of the data and to visualize the variables that could be extracted. We identified if data was presented in a structured or unstructured form, if there was missing data and if standardized vocabularies were utilized. Some records did not always include contact with the patient, and were related to a physician action in the record, or the receipt of a document (e.g., biology results). Therefore, we filtered on records with at least one mention of a drug prescription.

We described two populations to explore the database: the general population and the elderly aged 75 and more.

3. Results

3.1. Data integration

We implemented the tables PERSON, VISIT_OCCURRENCE, DRUG_EXPOSURE, MEASUREMENT, NOTE, OBSERVATION, LOCATION, PROVIDER in the OMOP-CDM. The following data quality problems were identified: missing birth date, structured data but without terminology, free text, non-standardized questions, missing zip codes and non-standardized city labels.

3.2. Descriptive analysis

Between 2013 and 2020, 16,396 patients were admitted for at least one consultation with a drug prescription. A total of 181,527 consultations were analyzed. The overall population had a mean (standard-deviation) age of 47.6 (24.2) and 54% were male, at consultation. The figure 1 displays the number of patients and consultation per age (at the consultation). There were 472 patients of 20 years old, who have benefited from 1,258 consultations, for a ratio of 2.66 consultations per patient, while there were 428 patients of 70 years old, who have benefited from 1.901 consultations, for a ratio of 4.44 consultations per patient. In average, 3.39 (2.88) drugs were prescribed per consultation for the general population. The regular measurement of biometric data during the consultations makes it possible to follow their evolution over time, as shown in figure 2 with the evolution of weight at different ages at the consultation.

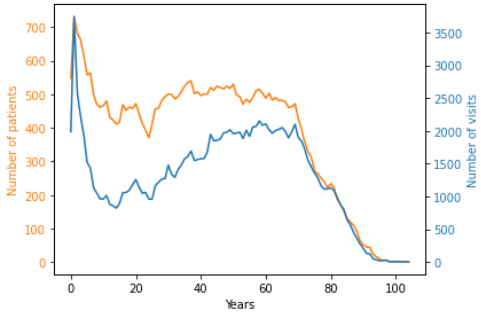


Figure 1. Number of patients and consultations per age at the consultation.

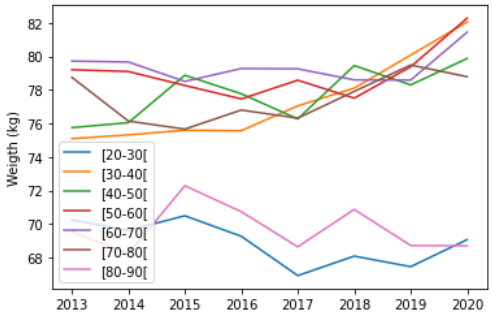


Figure 2. Change in weight from 2013 to 2020 for patients aged 20 (20 to 29), 30 (30 to 39), 40 (40 to 49), 50 (50 to 59), 60 (60 to 69), 70 (70 to 79), 80 (80 to 89) years old at consultation

3.3. Elderly

The elderly population (over 75 years of age) represents 5,5% of the general population or 900 patients for 13,867 consultations (7.6% of the consultations of the general population). This population received an average of 4.5 consultations per year. Elderly population benefited from 5.97 (3.72) drug prescriptions on average per consultation. The drugs prescribed corresponded to 871 unique ATC codes. The figure 3 represents the top 30 most prescribed drugs. The most prescribed drugs treat pain (paracetamol) for 10% of the prescriptions, followed by NSAIDs (acetylsalicylic acid) for 2.5% and esomeprazole for 2.3% of the prescriptions.

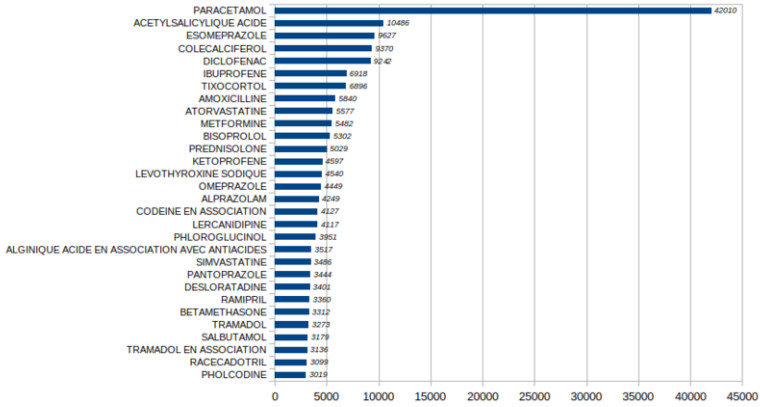


Figure 3. The top 30 most prescribed drugs for elderly patients.

4. Discussion

We integrated primary care data from a practice with 6 practitioners, over 8 years, for a total of 16,396 patients and 181,527 consultations with at least one prescription of drug. This allowed to highlight that the elderly population had more visits than the general population and received more medications. Surprisingly, the most prescribed drugs for elderly patients treat the pain rather than the symptomatology. With data collected in primary care, we can track clinical parameters over several years. The data comes directly from the source software and is more accurate than that collected secondarily for billing. Moreover, it contains blood tests results that are not available in national claim databases. This original nature of this type of data source offers the possibility of comparing the evolution of biology with treatment over the long term.

Despite the use of two standard terminologies (ICD10 and CIP), some data remain complicated to homogenize. In particular, the labels of the blood tests depend on the laboratory, and examinations and questionings are documented in free text during the consultation. Data quality problems in primary care EHR have already been identified in the literature [6,9] and we could apply the method proposed by Lacroix-Hugues et al. To map unstructured data (i.e., symptoms and diagnoses) to ICPC2 classification (International Classification of Reference for Primary Care) [6].

This work will be replicated in other centers equipped with the same software. By using the OMOP-CDM, we can also integrate data from other software editors for further collaborations in a multisite network.

5. Conclusions

We implemented a data warehouse with data collected from a GPs center. Despite many manuals' entries, we were able to produce longitudinal statistics. This work will be completed by integrating data from other GPs centers.

6. Funding

This research was funded by PreciDIAB, which is jointly supported by the French National Agency for Research (ANR- 18- IBHU- 0001), by the European Union (FEDER - Agreement NP0025517), by the Hauts-de-France Regional Council (Agreement 20001891/NP0025517) and by the European Metropolis of Lille (MEL, Agreement 2019_ESR_11).

References

[1] Schoen C, Osborn R, Squires D, Doty M, Rasmussen P, Pierson R, Applebaum S. A survey of primary care doctors in ten countries shows progress in use of health information technology, less in other areas. Health Aff (Millwood). 2012 Dec;31(12):2805-16. doi: 10.1377/hlthaff.2012.0884. Epub 2012 Nov 15. PMID: 23154997.

[2] Meystre SM, Lovis C, Bürkle T, Tognola G, Budrionis A, Lehmann CU. Clinical Data Reuse or Secondary Use: Current Status and Potential Future Progress. Yearb Med Inform. 2017 Aug;26(1):38-52. doi: 10.15265/IY-2017-007. Epub 2017 Sep 11. PMID: 28480475; PMCID: PMC6239225.

[3] Jannot AS, Zapletal E, Avillach P, Mamzer MF, Burgun A, Degoulet P. The Georges Pompidou University Hospital Clinical Data Warehouse: A 8-years follow-up experience. Int J Med Inform. 2017 Jun;102:21-28. doi: 10.1016/j.ijmedinf.2017.02.006. Epub 2017 Feb 16. PMID: 28495345.

[4] Moulis G, Lapeyre-Mestre M, Palmaro A, Pugnet G, Montastruc JL, Sailler L. French health insurance databases: What interest for medical research? Rev Med Interne. 2015 Jun;36(6):411-7. doi: 10.1016/j.revmed.2014.11.009. Epub 2014 Dec 26. PMID: 25547954.

[5] Gentil ML, Cuggia M, Fiquet L, Hagenbourger C, Le Berre T, Banâtre A, Renault E, Bouzille G, Chapron A. Factors influencing the development of primary care data collection projects from electronic health records: a systematic review of the literature. BMC Med Inform Decis Mak. 2017 Sep 25;17(1):139. doi: 10.1186/s12911-017-0538-x. PMID: 28946908; PMCID: PMC5613384.

[6] Lacroix-Hugues V, Darmon D, Pradier C, Staccini P. Creation of the First French Database in Primary Care Using the ICPC2: Feasibility Study. Stud Health Technol Inform. 2017;245:462-466. PMID: 29295137.

[7] de Lusignan S, van Weel C. The use of routinely collected computer data for research in primary care: opportunities and challenges. Fam Pract. 2006 Apr;23(2):253-63. doi: 10.1093/fampra/cmi106. Epub 2005 Dec 20. PMID: 16368704.

[8] Hripcsak G, Duke JD, Shah NH, Reich CG, Huser V, Schuemie MJ, Suchard MA, Park RW, Wong IC, Rijnbeek PR, van der Lei J, Pratt N, Norén GN, Li YC, Stang PE, Madigan D, Ryan PB. Observational Health Data Sciences and Informatics (OHDSI): Opportunities for Observational Researchers. Stud Health Technol Inform. 2015;216:574-8. PMID: 26262116; PMCID: PMC4815923.

[9] Terry AL, Stewart M, Cejic S, Marshall JN, de Lusignan S, Chesworth BM, Chevendra V, Maddocks H, Shadd J, Burge F, Thind A. A basic model for assessing primary health care electronic medical record data quality. BMC Med Inform Decis Mak. 2019 Feb 12;19(1):30. doi: 10.1186/s12911-019-0740-0. PMID: 30755205; PMCID: PMC6373085.

Challenges of Trustable AI and Added-Value on Health
B. Séroussi et al. (Eds.)

doi:10.3233/SHTI220511

Patients' Access to Their Psychiatric Records - A Comparison of Four Countries

Annika BÄRKÅS[a,1], Maria HÄGGLUND[a], Jonas MOLL[b], Åsa CAJANDER[c], Hanife REXHEPI[d], Iiris HÖRHAMMER[e], Charlotte BLEASE[f], Isabella SCANDURRA[b]
[a] *Department of Women's and Children's Health, Uppsala University, Sweden*
[b] *Informatics, School of Business Örebro University, Örebro, Sweden*
[c] *Department of Information Technology, Uppsala University, Sweden*
[d] *School of Informatics, Skövde University, Sweden*
[e] *Department of Industrial Engineering and Management, Aalto University, Finland*
[f] *General Medicine and Primary Care, Beth Israel Deaconess Medical Center, Harvard Medical School, Boston, MA, USA*

Abstract. Several Nordic and Baltic countries are forerunners in the digitalization of patient ehealth services and have since long implemented psychiatric records as parts of the ehealth services. There are country-specific differences in what clinical information is offered to patients concerning their online patient accessible psychiatric records. This study explores national differences in Sweden, Norway, Finland, and Estonia in patient access to their psychiatric records. Data was collected through a socio-technical data collection template developed during a workshop series and then analyzed in a cross-country comparison focusing on items related to psychiatry records online. The results show that psychiatric records online are offered to patients in all four countries, and provide the same functionality and similar psychiatry information. Overall, the conclusion is that experiences of various functionalities should be scrutinized to promote transparency of psychiatric records as part of the national eHealth services to increase equality of care and patient empowerment.

Keywords. mental health, psychiatry, psychiatric record, psychiatric notes, patient accessible electronic health record, PAEHR, open notes

1. Introduction

Patients who can access and read their psychiatric record online perceive an increased understanding of their mental health [1,2], a better awareness of potential side effects of, and adherence to, their medications [2], increased feeling of validation [3], and are feeling in control of their care [3]. Several studies have reported these results investigating mental health patients' experiences of reading their mental health notes online, the majority conducted in the USA [4].

Implementation of patient access to Electronic Health Records (EHR) online, also referred to as Patient Accessible Electronic Health Record (PAEHR), is becoming more

[1] Corresponding Author, Annika Bärkås; E-mail: annika.barkas@kbh.uu.se.

widespread internationally [5]. Via the PAEHR, patients can access and read notes written by clinicians, see lab and test results, diagnoses, and prescribed medications. The new federal law 21st Century Cures Act in the US, mandated all healthcare providers since April 2021 to offer patients access to their clinical information housed in their EHR [6,7]. In Sweden, one of the studied countries here, all citizens were provided access to the PAEHR in 2018, initially implemented in one healthcare region in 2012, and stepwise disseminated to all other 20 regions [7].

Despite positive patient experiences reading the psychiatric record online, access to clinical information from psychiatric care is considered particularly controversial. Studies investigating healthcare professionals' experiences of patients accessing their psychiatric records online report clinicians' concerns of patients becoming confused, anxious, or offended by what they read [2,8–12]. Swedish and US studies report that clinicians in psychiatric care would be pleased if psychiatric records online were discontinued [9,10]. An early Swedish study conducted in 2014, before patients could access their psychiatry records online, reports psychiatry clinicians' fear of increased threats and violence from patients if they would get access to their psychiatric records [13]. In addition, a Norwegian study reports that 29% of clinicians in psychiatric care hide parts of clinical information that should be presented in the PAEHR [14]. Several Nordic and Baltic countries are forerunners in digitalization of patient services and have since long implemented PAEHRs and psychiatric records as parts of the eHealth services. Implementation inevitably entails country-specific and regional differences in what clinical information is offered. The contribution of this paper is to increase the knowledge about differences in how online psychiatric records are implemented. We aim to explore national differences in Sweden, Norway, Finland, and Estonia concerning the functionality, content, and policies of the online psychiatric records. The study is conducted within the international research project NORDeHEALTH [15], which identifies challenges and opportunities in digitalization of health services for patients.

2. Methods

A socio-technical analysis of national patient portals was carried out in Sweden, Estonia, Finland, and Norway. The socio-technical framework proposed by Sittig and Singh [16], especially tailored to health systems, guided the data collection and the analysis. For each of the eight framework dimensions: *Hardware & Software Computing Infrastructure; Clinical Content; Human-Computer Interface; People; Workflow & Communication; Internal Organizational Policies, Procedures & Culture; External Rules, Regulations & Pressures; System Measurement & Monitoring*, as well as the added dimension *Features & Functions,* several questions were developed to cover socio-technical situations regarding development, implementation and use of patient portals. The first version of the data collection instrument was inspired by a recent socio-technical analysis [17], and the original data collection was carried out during four digital workshops with 1-4 topic experts from the above countries between June and November 2021. Between workshops, the topic experts had time to consult and verify the information with representatives from the national providers. A complementary, smaller data collection based on the dimension *Clinical Content* was carried out in December 2021 to elaborate on a subset regarding psychiatry records online. The digital workshop material consisted of a shared google excel sheet prepared with a list of identified questions. The topic experts from each country wrote down their answers in the shared google excel sheet.

The collected data were analyzed by comparing the answers from each country (Table 1). The analysis was based on these questions from the excel sheet:

1. What type of notes that regard psychiatry care does the PAEHR provide?
2. Does the PAEHR provide access to read prescribed psychiatric medications?
3. Does the PAEHR provide psychiatric diagnoses?
4. Are there policies that care professionals can lean on to limit certain patient groups to read their notes (e.g. if considered at suicidal risk)?

3. Results

Results answering questions 1-3 are presented in Table 1, and disclose that all countries offer similar amounts of clinical information from psychiatric outpatient and inpatient care through digital health services. Sweden, Estonia, and Finland offer patients access to their notes from all psychiatric care settings, such as adults, pediatric-adolescents, and forensic settings. All countries offer the similar amounts of information units presented in Table 1.

Table 1. Information units offered to patients in psychiatric care, comparing four national health services online; Journalen (Sweden), Digilugu (Estonia), Omakanta (Finland), and HelseNorge (Norway).

Information units	**Sweden** Journalen	**Estonia** Digilugu	**Finland** Omakanta	**Norway** HelseNorge
Type of notes that regard psychiatry care	Psychiatrists; Other physicians; Psychologists; Nurses; Social workers; Counselors; Referrals; *MDT notes	Discharge summaries; Referrals	Psychiatrists; Other physicians; Psychologists; Nurses; Discharge summaries; Referrals	Psychiatrists; Other physicians; Psychologists; Nurses; Social workers; Discharge summaries; MDT notes
Outpatient psychiatric care	Yes	Yes	Yes	Yes
Inpatient psychiatric care	Yes	Yes	Yes	Yes
Notes are shared in these psychiatric settings:				
Forensic	Yes	Yes	Yes	Yes
Adults	Yes	Yes	Yes	Yes
Pediatric-Adolescents	Yes	Yes	Yes	Yes
Access to read prescribed psychiatric medications	Yes	Yes	Yes	Yes
Psychiatric diagnoses	Yes	Yes	Yes	Yes

*MDT notes are notes written during Multidisciplinary team conferences.

Although all countries in this study give access to both out- and inpatient notes, there are differences regarding the types of notes patients can access.

Results from question 4 show that several safeguards were also used in cases where patients may put themselves or others at risk (e.g. suicidal risks or violence in close relationships). In Sweden, some regions have implemented specific templates for documenting, such as suicide risks, which are not shared with patients [18]. This also applies throughout Finland. Such templates are not used in Norway or Estonia. In both Sweden and Finland, embargo times are implemented in inpatient psychiatric care so that notes are not immediately accessible to patients. This functionality applies in one Norwegian region, while two other regions give immediate access (also in outpatient

psychiatric care). In Estonia, patients have immediate access to psychiatric notes but only signed notes. The latter also applies in Norway. In Sweden, patients can read both signed and unsigned psychiatric notes, and as a patient, one can filter which type one want to read. In Finland, patients only can access and read signed notes. Bärkås et al. (2021) explain the meaning of signed notes as a correct and complete note, validated by the clinician responsible for the information in the note. The responsible clinician has not yet confirmed or validated an unsigned note.

4. Discussion and Conclusion

This is the first comparison of PAEHR psychiatric information in Sweden, Estonia, Finland, and Norway. The results show that all four countries offer patients in psychiatric care access to read their psychiatric records online, however, with differences in the amounts of clinical information shared. In Sweden, Finland, and Norway, patients can read their notes from several healthcare professions in psychiatric care. In Estonia, psychiatry notes are limited to discharge summaries and referrals. On the other hand, the lack of delay period and policies limiting the access for any user in Estonia may compensate for how the user experiences the access. Further research is therefore needed to assess real usage by patients, as well as to cover the knowledge gap of patients' and clinicians' current experiences of psychiatric records online in the northern countries and beyond. Future work may also be to compare the implications of immediate or delayed access to clinical information and whether the notes are shared unsigned or signed. The in-depth investigation in [18] could be used as a point of departure.

One unintended consequence of patient access to psychiatric records may be a reduction of stigmatization of people with mental health conditions since psychiatric notes are shared in the same way as somatic notes. In addition, this could reduce psychiatry clinicians' perceived concerns about PAEHRs reported in previous studies [2,8–12] and apprehension of increased violence from patients reading their psychiatric records since no evidence supports the connection of increased violence and PAEHRs [13]. Previous studies almost exclusively report positive patient experiences of accessing psychiatric records [1–4], however, it is conceivable that some patients, for example, persons with borderline personality disorder, posttraumatic stress disorder, patients with eating disorders, or those with suicidal ideation, might become so deeply upset by what they read that access worsens symptoms [19,20]. Further research is needed to understand potential negative experiences of PAEHRs in mental health care [21].

This study concludes that patients receiving psychiatric care in Sweden, Finland, Norway, and Estonia are offered access to their psychiatric records online. Patients receiving psychiatric care in Estonia have access to fewer psychiatric information items in their PAEHR than in other countries, encouraging further studies related to perceived use. There is a call for homogeneity for the various PAEHR systems among the countries [5]. This study shows that these four countries, despite e.g. different systems and implementation models, provide the same functionality and similar psychiatry information to the patients. Our overall conclusion is that experiences of various functionalities should be fully scrutinized to explore the effects of transparency of psychiatric records, as parts of national eHealth services, including investigating the potential to increase care equity and patient empowerment.

Acknowledgement

This work was supported by NordForsk through the funding to Nordic eHealth for Patients: Benchmarking and Developing for the Future (NORDeHEALTH), project number 100477.

References

[1] Leung K, Clark C, Sakal M, Friesen M, Strudwick G. Patient and Family Member Readiness, Needs, and Perceptions of a Mental Health Patient Portal: A Mixed Methods Study. Stud Health Technol Inform. 2019;257:266–70.
[2] Blease C, Dong Z, Torous J, Walker J, Hägglund M, DesRoches CM. Association of Patients Reading Clinical Notes with Perception of Medication Adherence among Persons with Serious Mental Illness. JAMA Netw Open. 2021 Mar;4(3):e212823.
[3] O'Neill S, Chimowitz H, Leveille S, Walker J. Embracing the new age of transparency: mental health patients reading their psychotherapy notes online. J Ment Health. 2019 Oct;28(5):527–35.
[4] Schwarz J, Bärkås A, Blease C, Collins L, Hägglund M, Markham S, et al. Sharing Clinical Notes and Electronic Health Records With People Affected by Mental Health Conditions: Scoping Review. JMIR Ment Heal. 2021 Dec;8(12):e34170.
[5] Essén A, Scandurra I, Gerrits R, Humphrey G, Johansen MA, Kierkegaard P, et al. Patient Access to Electronic Health Records: Differences Across Ten Countries. Health Policy Technol. 2018 Mar;7(1):44–56.
[6] Blease C, Salmi L, Hägglund M, Wachenheim D, DesRoches C. COVID-19 and Open Notes: A New Method to Enhance Patient Safety and Trust. JMIR Ment Heal. 2021 Jun;8(6):e29314.
[7] Blease C, Salmi L, Rexhepi H, Hägglund M, DesRoches CM. Patients, clinicians and open notes: information blocking as a case of epistemic injustice. J Med Ethics. 2021 May;medethics-2021-107275.
[8] Petersson L, Erlingsdóttir G. Open Notes in Swedish Psychiatric Care (Part 1): Survey Among Psychiatric Care Professionals. JMIR Ment Heal. 2018 Feb;5(1):e11.
[9] Dobscha SK, Denneson LM, Jacobson LE, Williams HB, Cromer R, Woods S. VA mental health clinician experiences and attitudes toward OpenNotes. Gen Hosp Psychiatry. 2016 Jan-Feb;38:89–93.
[10] Petersson L, Erlingsdóttir G. Open Notes in Swedish Psychiatric Care (Part 2): Survey Among Psychiatric Care Professionals. JMIR Ment Heal. 2018 Jun;5(2):e10521.
[11] Strudwick G, Clark C, Sanches M, Strauss J. Predicators of Mental Health Professionals' Perceptions of Patient Portals. AMIA Symp Proc. 2018 Dec;989–97.
[12] Denneson LM, Cromer R, Williams HB, Pisciotta M, Dobscha SK. A Qualitative Analysis of How Online Access to Mental Health Notes Is Changing Clinician Perceptions of Power and Therapeutic Relationship. J Med Internet Res. 2017Jun;19(6):e208.
[13] Åkerstedt U, Cajander Å, Moll J, Ålander T. On Threats and Violence for Staff and Patient Accessible Electronic Health Records. Cogent Psychol. 2018 Sep;5(1):1–11.
[14] Kristiansen E, Johansen MA, Zanaboni P. Healthcare personnels' experience with patients' online access to electronic health records: Differences between professions, regions, and somatic and psychiatric healthcare. SHI 2019. Proceedings of the 17th Scandinavian Conference on Health Informatics. November 12-13, 2019; Oslo, Norway. Linköping Electronic Conference Proceedings 161:16. p. 93–8.
[15] Hägglund M. Nordic countries lead new initiative on patient access to EHRs. 2021 May 18 [cited 2022 Jan 21]. In: BMJ (International Ed.). Available from: https://blogs.bmj.com/bmj/2021/05/18/maria-hagglund-nordic-countries-lead-new-initiative-on-patient-access-to-ehrs/
[16] Sittig DF, Sing H. A new socio-technical model for studying health information technology in complex adaptive healthcare systems. Qual Saf Health Care. 2010 Oct;19 Suppl 3(Suppl 3):i68–74.
[17] Hägglund M, Scandurra I. A Socio-Technical Analysis of Patient Accessible Electronic Health Records. Stud Health Technol Inform. 2017;244:3–7.
[18] Bärkås A, Scandurra I, Rexhepi H, Blease C, Cajander Å, Hägglund M. Patients' Access to Their Psychiatric Notes: Current Policies and Practices in Sweden. Int J Environ Res Public Health. 2021 Aug;18(17):9140.
[19] Strudwick G, Yeung A, Gratzer D. Easy Access, Difficult Consequences? Providing Psychiatric Patients With Access to Their Health Records Electronically. Front Psychiatry. 2019 Dec;10:917.
[20] Blease C, O'Neill S, Walker J, Hägglund M, Torous J. Sharing notes with mental health patients: balancing risks with respect. Lancet Psychiatry. 2020 Nov;7(11):924–5.
[21] Blease C, O'Neill SF, Torous J, DesRoches CM, Hägglund M. Patient Access to Mental Health Notes: Motivating Evidence-Informed Ethical Guidelines. J Nerv Ments Dis. 2021 Apr;209(4):265–9.

Challenges of Trustable AI and Added-Value on Health
B. Séroussi et al. (Eds.)

doi:10.3233/SHTI220512

Human Factors of EHR Adoption in Saudi Primary Healthcare

Bander Dowahi ALANAZI[a,1] , Mohammed Hijji ALHIJJI[b] and Ibtisam Dhuwayhi ALANAZI[c]
[a] *College of Health and Medicine, University of Tasmania, Launceston, Tasmania, Australia*
[b] *College of Dentistry, King Saud University, Riyadh, Saudi Arabia*
[c] *Accounting Department, Prince Sattam bin Abdulaziz University, Riyadh, Saudi Arabia*

Abstract. Electronic Health Records are rapidly gaining traction in healthcare with increased acceptance and adoption. However, there is limited understanding of factors influencing adoption in primary care. This paper investigates the human factors of EHR adoption in primary healthcare in Saudi Arabia. An online survey questionnaire was sent to all primary healthcare professionals in Riyadh city, Saudi Arabia. A 65.9% (1127/1710) response rate was obtained. The respondents demonstrated positive perceptions of EHRs in relation to the systems' benefits. The perceptions were influenced by sociodemographic variables; hence, need consideration when implementing EHRs in primary care.

Keywords. EHRs, healthcare professionals, perceptions, primary care

1. Introduction

A growing body of literature suggests that Electronic Health Records (EHRs) are increasingly adopted at different levels of care, including primary healthcare due to their potential benefits, such as improved work efficiency, better quality of health care, and reduced medical costs [1-3]. However, despite the efforts to deploy EHRs in healthcare, implementation challenges still abound. World Health Organization (WHO) note that the adoption rates are still much lower, especially in the low- and middle-income countries, due to several factors [4]. Health care providers' perceptions, attitudes, intentions, and behaviors have been found to influence the adoption and use of EHRs [5, 6]. These tendencies are also affected by a range of factors related to individual, organizational, and system characteristics [5-10].

In Saudi Arabia, the ongoing healthcare system reforms have led to the introduction of EHRs at the primary care level, with all primary care centers (PCCs) currently transitioning from paper-based records to electronic systems [8]. Few studies on this area show that health care providers' perceptions are critical determinants of EHR adoption in Saudi hospitals [8, 11, 12]. Individual and system attributes have also been identified to influence healthcare professionals' perceptions in hospitals in the GCC region [13].

[1] Bander Dowahi Alanazi, College of Health Science, University of Tasmania, Launceston, Tasmania, Australia; E-mail: bander12@hotmail.com

However, little is known about the influencing factors at the primary care level. This paper examines the potential role of sociodemographic factors on perceptions and adoption of EHRs in primary care in Saudi Arabia.

2. Methods

All (1710) healthcare professionals working in PCCs in Riyadh city, Saudi Arabia, and are directly interacting with EHRs to provide care were invited to participate in a large survey evaluating perceptions about the adoption of EHRs [14]. They comprised of physicians, nurses, pharmacists, technicians, and allied health professionals. The survey questionnaire was adapted from a previous study in Turkey [15]. It had three components related to perceived benefits of EHR adoption (14 items), obstacles to EHR adoption (9 items), and satisfaction with EHRs (10 items) in primary care. The tool was deployed online using Research Electronic Data Capture (REDCap) between November 30, 2017 and January 30, 2018. University of Tasmania Social Science Human Research Ethics Committee and the Ministry of Health of Saudi Arabia approved the study.

3. Results

3.1. Sociodemographic characteristics of the participants

The completed surveys were 1127, a 65.9% response rate. Most of the respondents were nurses, females, Saudis, aged 20 – 34 years, had less than 10 years of work experience in primary care, and lacked previous health experience outside the Kingdom of Saudi Arabia (KSA), previous EHR training, and experience with EHR. These sociodemographic characteristics of the respondents are presented in Table 1.

Table 1. Sociodemographic characteristics of the participants

Demographic characteristics of participants		Respondents (N=1127)
Occupation	Physician	209 (18.5%)
	Nurse	367 (32.6%)
	Pharmacist	208 (18.5%)
	Technician	228 (20.2%)
	Other	115 (10.2%)
Gender	Male	503 (44.6%)
	Female	624 (55.4%)
Nationality	Saudi	811 (72.0%)
	Non-Saudi	316 (28.0%)
Age (years)	20–34	608 (53.9%)
	35–49	471 (41.8%)
	50+	48 (4.3%)
Length of work experience in primary healthcare (years)	0–10	870 (77.2%)
	11–20	226 (20.1%)
	21+	31 (2.8%)
Previous health experience outside KSA	No	686 (60.9%)
	Yes	441 (39.1%)
Previous training in EHRs in primary healthcare	No	674 (59.8%)
	Yes	453 (40.2%)
Previous EHR experience in primary healthcare	No	686 (60.9%)
	Yes	441 (39.1%)

3.2. Perceptions of benefits, obstacles and satisfaction with EHRs

The respondents indicated a high agreement with all the statements related to the benefits and satisfaction with EHR. In contrast, all the perceived obstacles had a low agreement level, except for the statement that an EHR system 'needs frequent revisions due to technological developments that had an agreement level of 45.3% (Table 2).

Table 2. Agreement levels with perceived benefits, obstacles and satisfaction items

Dimension	Item	Level of agreement (%)
Benefits of adopting EHR in primary care	Provides quick and reliable access to scientific research	76.0
	Enables easy access to information from past medical records	73.9
	Provides access to patient data and analysis	76.8
	Provides better data	74.5
	Makes it easy to transfer data	74.4
	Provides access to practice standards	73.6
	Enables following test results	75.3
	Saves time in documenting health data	75.2
	Decreases paper-based documentation	77.1
	Improves the feeling of professionalism	76.2
	Improves communication between health professionals and patients	73.7
	Contributes to health professionals' ability to make patient care decisions	72.3
	Improves communication between health professionals	73.5
	Reduces medical errors	63.5
Obstacles to adopting EHRs in primary care	Is too complicated and not user-friendly	17.0
	Compromises patient safety	18.8
	Decreases interaction between the health professional and patient	20.5
	Increases health professionals' workloads	23.5
	It is difficult to provide data security in EHRs	19.6
	Consumes more time than paper-based systems	25.2
	Is 'down' frequently	20.9
	Is costly	24.8
	Needs frequent revisions related to technological developments	45.3
Satisfaction with EHR in primary care	I feel EHR is useful	78.9
	I feel EHR is an important system for primary health care centers	77.0
	I feel EHR has been successful in primary health care centers	67.7
	I feel EHR is worth the time and effort required to use it	70.1
	I feel the EHR improves the quality of healthcare services in primary healthcare centers	74.8
	I feel the quality of my work has improved	68.2
	I feel the quality of information has improved due to EHR	70.9
	I feel my performance has improved due to EHR	65.9
	I feel patient safety has improved due to EHR	65.8
	Overall, I am satisfied with the EHR system in primary healthcare centers	69.1

3.3. Relationships between sociodemographic characteristics and perceptions of benefits, obstacles and satisfaction with EHRs

The healthcare professionals' perceptions varied with sociodemographic characteristics. There were significant differences across all perceived benefits in relation to occupation, age, experience outside KSA, and EHR training ($p<0.05$). For example, physicians, older respondents (≥50 years), and those without experience outside the KSA indicated the

highest agreement with all the benefits. Gender, nationality, and work experience also demonstrated significant differences across most benefits. In contrast, there were no significant differences across all benefits except for 'improves communication between health professionals' based on previous EHR experience.

There were significant differences across all items for obstacles by occupation and age. Nationality and experience outside KSA were associated with significant relationships in perceptions of most obstacles. However, non-Saudis indicated a lower agreement with all the obstacle items. There were also no significant differences across all the obstacle items with respect to previous EHR experience.

Lastly, there were significant differences across all satisfaction items in relation to occupation, nationality, age, length of work experience, experience outside KSA, and EHR training. For example, physicians and respondents with no experience outside the KSA and EHR training had a higher agreement with all the items than their counterparts. Non-Saudis and respondents aged 50 years and above, and with >20 years of experience had a higher agreement level with most items compared to professionals who are Saudis, younger and have fewer years of experience. In addition, most satisfaction items showed significant differences across gender but not in terms of previous EHR experience.

4. Discussion

This study reported a high agreement with statements for benefits and satisfaction with EHRs but a low one with obstacles. There were significant differences in respondents' perceptions of EHR benefits based on occupation, age, previous experience working outside KSA, and previous training in EHRs ($p<0.05$). The perceptions were significantly associated with only occupation and age for the obstacles to EHR adoption. However, satisfaction with EHRs was significantly associated with all the demographic variables except previous experience using EHR in primary care.

Similar to the results of previous studies [7, 15, 16], the findings suggest that the respondents generally had a positive perception of EHRs and were more likely to accept and use the systems in primary care. However, they associations with sociodemographic variables indicate that personal attributes also affect EHR adoption. The significant differences across the sociodemographic factors except for previous experience in EHR suggest that the perception and adoption of EHRs are likely to vary among health care providers. For example, physicians tend to have a more positive perception and are more likely to adopt EHRs compared to other health care providers, as demonstrated by the highest agreement level with benefit and satisfaction items. Other studies have also supported these findings [16, 17]. Of great concern is the finding that Saudi healthcare professionals appear to have a less positive perception of EHRs compared to non-Saudis. Although the finding could be attributed to prior exposure and more knowledge about EHR systems among non-Saudi healthcare professionals, it signifies challenges that the policymakers may face with implementation of the Saudisation policy that aims to increase the Saudi workforce in all sectors, including healthcare. Training of Saudi care providers in EHRs could be crucial in improving adoption in this setting as training has been shown to improve knowledge and skills about the EHR system as well as increasing health care providers' confidence, satisfaction and positive perception [7, 18].

Overall, this study found a positive perception and satisfaction with the EHRs among healthcare professionals in PCCs in Saudi Arabia. It, however, highlighted the impact that sociodemographic variables could have on the adoption of these electronic systems

intended to revolutionize the delivery and provision of health care services. The study also shows that these factors should be taken into account when implementing EHRs in healthcare settings. Thus, this study makes a unique contribution to knowledge about the role of human factors in the adoption of EHRs in Saudi primary care and similar settings. Future research should investigate a full spectrum of factors that may influence healthcare professionals' perceptions in order to inform the system's adoption and implementation.

References

[1] Manca DP. Do electronic medical records improve quality of care?: Yes. Canadian Family Physician. 2015;61(10):846-7.

[2] Hoover R. Benefits of using an electronic health record. Nursing2018. 2016;46(7):21-2.

[3] Menachemi N, Collum TH. Benefits and drawbacks of electronic health record systems. Risk management and healthcare policy. 2011;4:47.

[4] Organization WH. Global diffusion of eHealth: making universal health coverage achievable: report of the third global survey on eHealth: World Health Organization; 2017.

[5] Bahadori M, Alimohammadzadeh K, Abdolkarimi K, Ravangard R. Factors Affecting Physicians' Attitudes Towards the Implementation of Electronic Health Records Using Structural Equation Modeling Modeling (SEM). Shiraz E-Medical Journal. 2017;18(11).

[6] Mathai N, McGill T, Toohey D. Factors influencing consumer adoption of electronic health records. Journal of Computer Information Systems. 2020:1-11.

[7] O'Donnell A, Kaner E, Shaw C, Haighton C. Primary care physicians' attitudes to the adoption of electronic medical records: a systematic review and evidence synthesis using the clinical adoption framework. BMC medical informatics and decision making. 2018;18(1):1-16.

[8] AlJarullah A, Crowder R, Wald M, Wills G. Factors Affecting the Adoption of EHRs by Primary Healthcare Physicians in the Kingdom of Saudi Arabia: An Integrated Theoretical Framework. 2018.

[9] Tubaishat A. Perceived usefulness and perceived ease of use of electronic health records among nurses: Application of Technology Acceptance Model. Informatics for health & social care. 2018;43(4):379-89. Epub 2017/09/19. doi: 10.1080/17538157.2017.1363761. PubMed PMID: 28920708.

[10] Gagnon M-P, Simonyan D, Godin G, Labrecque M, Ouimet M, Rousseau M. Factors influencing electronic health record adoption by physicians: A multilevel analysis. International Journal of Information Management. 2016;36(3):258-70.

[11] Alanezi F. Factors affecting the adoption of e-health system in the Kingdom of Saudi Arabia. International Health. 2020.

[12] Alsulame K, Khalifa M, Househ M. E-Health status in Saudi Arabia: A review of current literature. Health Policy and Technology. 2016;5(2):204-10.

[13] Alanazi B, Butler-Henderson K, Alanazi MR. Factors Influencing Healthcare Professionals' Perception towards EHR/EMR Systems in Gulf Cooperation Council Countries: A Systematic Review. Oman medical journal. 2020;35(6):e192.

[14] Alanazi B. Evaluating the healthcare professionals' perceptions about the adoption of electronic health records in primary care centres in Riyadh City, Saudi Arabia: University of Tasmania; 2020

[15] Secginli S, Erdogan S, Monsen KA. Attitudes of health professionals towards electronic health records in primary health care settings: a questionnaire survey. Informatics for Health and Social Care. 2014;39(1):15-32.

[16] Bani-issa W, Al Yateem N, Al Makhzoomy IK, Ibrahim A. Satisfaction of health-care providers with electronic health records and perceived barriers to its implementation in the United Arab Emirates. International journal of nursing practice. 2016;22(4):408-16.

[17] Kim S, Lee K-H, Hwang H, Yoo S. Analysis of the factors influencing healthcare professionals' adoption of mobile electronic medical record (EMR) using the unified theory of acceptance and use of technology (UTAUT) in a tertiary hospital. BMC medical informatics and decision making. 2015;16(1):1-12.

[18] Al-Harbi A. Healthcare providers' perceptions towards health information applications at King Abdul-Aziz Medical City, Saudi Arabia. International Journal of Advanced Computer Science and Applications.2011;2(10):10-31.

Challenges of Trustable AI and Added-Value on Health
B. Séroussi et al. (Eds.)

doi:10.3233/SHTI220513

Assessment of Health Service Quality Through Electronic Health Record – A Scoping Review

Hanna von GERICH[a,b1] and Laura-Maria PELTONEN[b]
[a] *Turku University Hospital, Turku, Finland*
[b] *Department of Nursing Science, University of Turku, Turku, Finland*

Abstract. The World Health Organization defines, that high quality health services should be effective, safe, people-centered, timely, equitable, integrated, and effective. This requires systematic quality assessment. The aim of this scoping review was to explore how electronic health records (EHRs) have been used to assess quality of health services using the WHO criteria. A total of 4247 records were obtained whereof 8 studies were included in the review. Research showed that EHRs were used to evaluate safety, performance and care processes. EHRs were regarded as an applicable real-world data source, highlighting the importance of consistency and standardised terminologies. Use of EHR data is limited to its representation of the real world and current evaluation systems have limited quality criteria, diverse definitions and they use only structured data. Future research should explore possibilities of natural language processing methods and include narrative EHR information for a more a comprehensive view of service quality assessment.

Keywords. Care quality, EHR, health service research, quality assessment

1. Introduction

The World Health Organization (WHO) defines quality of care as *"the degree to which health services for individuals and populations increase the likelihood of desired health outcomes"*. Health services should be effective, safe, people-centered, timely, equitable, integrated and efficient to meet expected standards [1]. However, determining the quality of health services is a multifaceted task, which requires validated instruments and systematic measurement, as well as systems thinking [2]. Furthermore, the service quality needs to be evaluated from different perspectives on individual, departmental, organisational, national and international levels.

Health care system quality of care may traditionally be explored through the Donabedian model, which divides essential elements in structures, processes and outcomes, where the structure includes the context of care delivery, processes the actions taken in care provision, and outcomes the impact of provided care on individuals [3]. It is the responsibility of the care provider to ensure the highest possible standard of services based on service users' needs and monitoring delivered care becomes key in quality improvement. The WHO suggests a process for building a strategy for quality,

[1] Corresponding Author, Hanna von Gerich, Department of Nursing Science 20014 University of Turku, Turku, Finland.; E-mail: hanna.m.vongerich@utu.fi.

which includes seven elements related to an initial analysis, a strategy development, and an implementation phase [4].

Healthcare digitalisation provides an opportunity for semiautomated comprehensive health service quality evaluation based on large and diverse data sets. National, regional and organisational databases are used to monitor different aspects of quality in health care [5]. Electronic health records (EHR), containing multidisciplinary documentation from the point of care may be seen as vast collections of holistic key data regarding individuals and their care within an organisation. Hence, EHRs provide an amazing and almost real-time opportunity for quality assessment and improvement. However, a recent study in the USA showed that there are concerns about EHR-based quality measures related to data access, standardisation of data elements and cost regarding integration between different systems [6]. Hence, the aim of this scoping review was to explore how EHRs have been used to assess health service quality, defined by the WHO criteria [1].

2. Methods

This scoping review was undertaken following Prisma-ScR Cheklist [7] and the framework by Arksey and O'Malley [9]. The following research question were identified:

1) What health service quality related factors have been assessed and how from EHRs?
2) What strengths and weaknesses are associated with utilising EHR data for assessing health service quality?
3) What are the next steps in utilising EHRs in health service quality assessment as reported by researchers?

2.1. Identifying Relevant Studies

For this scoping review, the PubMed, CINAHL (Ebsco) and Embase databases were searched using the search phrase: ((effective* OR safe* OR People-cent* OR timel* OR Equitab* OR Integrat* OR efficien*) AND ("health service") AND (EHR OR MHR OR note* OR record*)). Applicable MeSH terms including "Electronic Health Records"[Mesh] and "Medical Records"[Mesh] in PubMed and Subject headings (MH "Electronic Health Records+"), (MH "Medical Records+"), (MH "Patient Record Systems+"), (MH "Nursing Records") and (MH "Documentation+") in CINAHL were added to the searches. The search was done in January 2022. No time limits were used.

2.2. Study Selection

The studies included in this review were peer-reviewed scientific articles written in English, assessing the applicability of one or more WHO criteria for quality health services in quality evaluation using electronic health records. We excluded studies that were not relevant to nursing, manual records, research protocols and literature reviews.

2.3. Charting the Data and Collating, Summarising and Reporting the Results

Studies were downloaded to the Rayyan web-tool for systematic reviews (rayaan.ai), where duplicates were removed, and title and abstract screening was performed. The screening was performed by two researchers, with uncertainties discussed together. The following data were extracted: Author(s), publication details, study location, aim of the study, method of study and data extraction, analysed data, scientific disciplines involved in the study, strengths and limitations in utilising EHR data and next steps in utilising EHR in health service quality assessment. Extracted data were collected to a spreadsheet and analysed following the principles of content analysis [9].

3. Results

A total of 4247 records were obtained, with PubMed providing 2029, CINAHL 1296 and Embase 922 results. After the removal of 756 duplicates, 3491 articles' titles and were screened, followed by abstract screening of 380 records. Altogether 31 full text articles were assessed for eligibility, resulting in 8 descriptive studies being included in the review. The studies were conducted between 2010 and 2021 in Great Britain [10-14], Australia [15-16], China [17] and Germany [12]. Scientific disciplines involved in the studies were nursing sciences [10,13,15-16], medical sciences [11-12,14,17], economics [10,12] and social sciences [16]. Four of the studies utilised EHRs to identify adverse events [10-12,14], two to evaluate safety, quality or performance of care [18,17] and two to review care processes [13,16].

The study designs were majorly quantitative, non-randomized studies [10-14,16-17] with one exception of a mixed method study design [15]. Assessments were made analysing standardised data derived from EHRs such as number of patient contacts, used care interventions or physical and mental health measurements [13-16], diagnostic entries or codes [10,12,17] or Read codes [11,14]. Data extraction method was specified in five studies, with three using automated data extraction methods [10-11,14] and two manual methods [15-16].

EHRs were unanimously regarded as an applicable data source, highlighting the importance of the consistency and nature of datasets in assessing service quality. They provided a valuable snapshot into real-world patient care [13,15], and when using standardised diagnostic coding as a quality measure, the results were regarded as highly objective [17]. However, small samples [15] and incomprehensive datasets [10-11,13-14,16] were perceived as a limitation, resulting in imprecise or incomplete quality measurement. Additionally, entries made in the records did not necessarily indicate the quality of care functions, only their existence [16]. Using more complex and diverse quality criteria [11] and having universal definitions [12] could have resulted in a more comprehensive analysis of service quality.

Six studies provided suggestions to help advance the usability and research of EHRs in assessing quality of services. Foremost, to provide best possible information to both researchers and clinicians, systemic changes in use of EHRs are needed in regards to what and how care measures, patient outcomes or adverse events are reported [12-14,16]. Additionally, the association between adverse events identified in the health records and nurse staffing levels, for example, could be further explored, advancing the impact of quality care [10]. Moreover, using data mining techniques with large datasets, and seeking unexplored patterns adjacent to adverse events could be intensified [11].

4. Discussion

The results show that use of EHR data for systematic evaluation of health service quality is still very limited, although it is on the increase. It is important to acknowledge that data documented in EHRs are diverse in quality [18] and this may impact evaluation outcomes. Therefore, it is important to improve documentation practices and evaluate documentation quality on a regular basis when used to evaluate services. EHR-based quality evaluation results may then be added to a broader set of indicators collected from a variety of sources such as patient surveys and different data bases. Potential benefits of EHRs as part of service quality evaluation have been reported previously as well [19].

Secondly, current approaches using EHR data for evaluating service quality focus on structured data alone without other advanced analytical methods, although documented care is largely in narrative format containing rich descriptions of the service users' health related issues, care process and provided care [20]. Therefore, future research should explore possibilities of using natural language processing and text mining of narrative information in EHRs. Combining structured and unstructured data as sources for quality assessment provides a more holistic view of the care documented.

Thirdly, included studies involved only either nursing or medical researchers and no study involved a combination of health experts from different domains, highlighting a clear lack of systems thinking and a need to increase engagement and representation of different health care professionals in the development process of such systems for a balanced and comprehensive approach to sustainable assessment of health service quality.

Finally, the findings indicate that means and definitions on evaluating quality of health services still differ. Previous research has also encountered this issue, as illustrated by a systematic review on health care quality measurement using Service Provision Assessment datasets. A sample of 34 studies indicated that quality constructs were operationalized in extremely different ways, vastly limiting the generalizability of the results. [21] Current systems do not support such tasks sufficiently due to e.g. access, standardisation and interoperability issues [6]. More work is needed to develop a comprehensive and balanced framework of standardised outcome measures to be used in (semi)automated quality assessment tools on different levels in the health system. This would support benchmarking of service quality both within and beyond organisations and countries.

Our study limitations include a limited number of databases searched and a lack of quality assessment of studies included in the review. Future research is needed to explore to what extent EHRs can be used (i.e. content, extent, validity, technical and practical perspectives) to contribute to an overall assessment of health service quality when combined with data from multiple other sources.

5. Conclusions

Use of EHR data in assessment of health service quality is still limited. Currently, structured data are used to assess patient safety, processes and performance -related issues. Development of a comprehensive framework with standardised indicators with an interdisciplinary effort has the potential to support appropriate use of advanced analytical methods for better use of the unexplored potential of EHR data in comprehensive, transferable and automatised evaluation of health service quality.

References

[1] World Health Organization, Organization for Economic Co-operation and Development, World Bank. Delivering quality health services: a global imperative for universal health coverage. WHO; 2018.
[2] National Academies of Sciences, Engineering, and Medicine. Crossing the Global Quality Chasm: Improving Health Care Worldwide. Washington (DC): National Academies Press (US); 2018.
[3] Donabedian A. The quality of care: How can it be assessed? JAMA. 1988;260(12):1743–8.
[4] World Health Organization. Quality of care: a process for making strategic choices in health systems. WHO; 2006.
[5] Wang Y, Kung K, Byrd TA. Big data analytics: Understanding its capabilities and potential benefits for healthcare organizations. Technol Forecast Soc Change. 2018 Jan;126:3-13.
[6] Perloff J, Sobul S. Use of electronic health record systems in accountable care organizations. Am J Manag Care. 2022 Jan;28(1):e31-4.
[7] Tricco AC, Lillie E, Zarin W, O'Brien KK, Colquhoun H, Levac D, et al. PRISMA extension for scoping reviews (PRISMA-ScR): checklist and explanation. Ann Intern Med. 2018;169(7):467-73.
[8] Arksey H, O'Malley L. Scoping studies: towards a methodological framework. Int J Soc Res Meth. 2005;8(1):19-32.
[9] Elo S, Kyngäs H. The qualitative content analysis process. J Adv Nurs. 2007;62(1):107-15.
[10] Murphy A, Griffiths P, Duffield C, Brady NM, Scott AP, Ball J, et al. Estimating the economic cost of nurse sensitive adverse events amongst patients in medical and surgical settings. J Adv Nurs. 2021 Aug;77(8):3379-88.
[11] Tsang C, Majeed A, Banarsee R, Gnani S, Aylin P. Recording of adverse events in English general practice: analysis of data from electronic patient records. Inform Prim Care. 2010 Jun;18(2):117-24.
[12] Friebel R, Henschke C, Maynou L. Comparing the dangers of a stay in English and German hospitals for high-need patients. Health Serv Res. 2021;56(Suppl. 3):1405-17.
[13] Guest JF, Fuller GW, Edwards J. Cohort study evaluating management of burns in the community in clinical practice in the UK: Costs and outcomes. BMJ Open. 2020 Jan;10(4):e035345.
[14] Tsang C, Bottle A, Majeed A, Aylin P. Adverse events recorded in English primary care: Observational study using the General Practice Research Database. Br J Gen Pract. 2013 Jan;63(613),e534-42.
[15] Gardner G, Gardner A, O'Connell J. Using the Donabedian framework to examine the quality and safety of nursing service innovation. J Clin Nurs. 2014 Jan;23(1):145-55.
[16] Lawn S, Zabeen S, Rowlands N, Picot S. Hidden care: Revelations of a case-note audit of physical health care in a community mental health service. Int J Ment Health Nurs. 2018 Dec;27(6):1742-55.
[17] Ji X, Fang Y, Liu J. Performance assessment of the inpatient medical services of a clinical subspecialty: A case study with risk adjustment based on diagnosis-related groups in China. Medicine. 2018 Jun;97(24):e10855.
[18] Akhu-Zaheva L, Al-Maaitah R, Bany Hani S. Quality of nursing documentation: Paper-based health records versus electronic-based health records: J Clin Nurs. 2018 Feb;27(3-4):e578-89.
[19] Shapiro LM, Kamal RN. Implementation of electronic health records during global outreach: A necessary next step in measuring and improving quality of care. J Hand Surg Am. 2022 Mar;47(3):279-83.
[20] Jefferies D, Johnson M, Griffiths R. A meta-study of the essentials of quality nursing documentation. Int J Nurs Pract. Apr 2010;16(2):112-24.
[21] Moucheraud C, McBride K. Variability in Health Care Quality Measurement among Studies Using Service Provision Assessment Data from Low- and Middle-Income Countries: A Systematic Review. Am J Trop Med Hyg. 2020 Sep;103(3),986-92.

Challenges of Trustable AI and Added-Value on Health
B. Séroussi et al. (Eds.)

doi:10.3233/SHTI220514

Patient Data Sharing and Reduction of Overtime Work of Nurses by Innovation of Nursing Records Using Structured Clinical Knowledge

Satoko TSURU[a,1], Tetsuro TAMAMOTO[b], Akihiro NAKAO[c], Kouichi TANIZAKI[a] and Naohisa YAHAGI[a,d]
[a] *The University of Tokyo, Japan*
[b] *Nara Medical University Hospital, Japan*
[c] *DoctorsMobile Corp., Japan*
[d] *Keio University, Japan*

Abstract. Half of nurses' overtime hours are due to records. Nursing records, which are mainly narrative records, cost a large amount of money. However, it has been pointed out that there are problems with their quality and post-use. In this study, we analyzed the value of nursing records for physicians. As a result, we found that the use of standard observation terms in nursing records can create an environment in which patients' conditions can be shared. To create this environment, the physicians of the clinical path committee classified hospitalized patients in terms of disease, treatment, and examination, and created a list of 778 process paths. Physicians, nurses, and researchers collaborated to develop digital contents with high-priority observation items and care actions adapted to patient conditions for each path. We developed a clinical support system equipped with these digital contents. In May 2019, we installed the system in a 900-bed university hospital. Then, in October 2020, we installed the system in a 400-bed general hospital. We used "nurses' overtime hours for recording" and "reduction rate" as indicators of the usefulness of this system. In the 900-bed university hospital, we compared the previous year's results for March, the end of the fiscal year. This overtime hours were 2,944 hours 00 minutes in March 2019 and 2,141 hours 55 minutes in March 2020. 27% reduction was indicated. The respective bed occupancy rates were 90.80 percent and 90.60 percent, with no difference. In the 400-bed general hospital, This overtime hours were compared to the previous year, covering November and December after one month of implementation. 386 hours in November 2019 and 204.5 hours in November 2020. 47% reduction indicated. 366 hours in December 2019 and 214.5 hours in December 2020. A reduction of 41% was shown. These results suggest that the implementation of this system can both improve the quality of team care and reduce overtime.

Keywords. quality management, overtime work, structured knowledge

[1] Corresponding Author, Satoko TSURU, The University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo, 113-8656 JAPAN ; E-mail: tsuru@tqm.t.u-tokyo.ac.jp

1. Introduction

Medical care is an intangible service. Therefore, without records, the medical care performed cannot be shared with others, nor can it be proven that it was performed. However, the recording process is time-consuming and labor-intensive. In Japan, the issue of overtime work by nurses has been emphasized [1-3]. Half of the overtime work of nurses is due to record making [4]. It has been pointed out that nursing records, which are mainly narrative, cost a great deal, but their quality is low, and they are difficult to use later [5,6]. The problems of nursing records can be discussed in terms of cost, quality, and conventions; however, they are difficult to solve and affect overtime work.

Tsuru et al. identified three key perspectives that make nursing records valuable to physicians. These were "symptoms of illness," "complications from surgical interventions," and "adverse events from medications" [7]. Tsuru et al. have been working to develop digital content that includes "observation of their occurrence" and "care to prevent their occurrence" as a standard nursing plan [8]. Fatigue from overtime has been reported to reduce the safety and quality of health care[9]. In addition, record omissions and delays in sharing records are problems directly related to medical safety. To address the complex related issues, we have realized a DX to improve the efficiency and quality of nursing planning and recording.

A clinical support system with 778 process paths was developed and introduced to a university hospital and general hospital. The purpose of this study was to compare overtime hours for recording nursing care before and after the introduction of the system at the two hospitals, and to determine the usefulness of this system.

2. Methods

This study was conducted in the following steps.

Step 1 (Process Path Design)

Four types of process paths, "Surgery," "Internal Medicine," and "Short-Term Hospitalization," which are generalized contents, and "COVID-19," which is an individualized content, were created using the Patient Condition Adaptive Path System (PCAPS) [9]. PCAPS has a model called "Clinical Process Chart" that visualizes the clinical process by connecting units representing the patient's condition. (Figure 1)

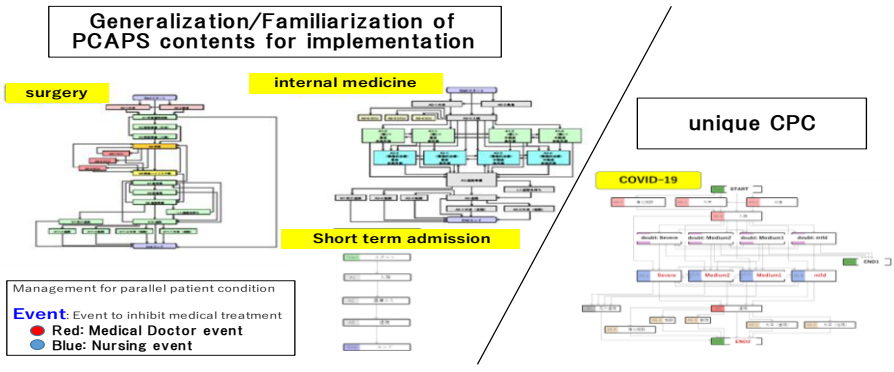


Figure 1. PCAPS Contents

Step 2 (Development of digital path content)
Physicians and surgeons on the Clinical Path Committee of a 900-bed university hospital identified the types of path content needed by each department using the perspectives of disease, treatment, and examination. As a result, 778 clinical process charts (CPCs) were identified. Physicians, nurses, and researchers collaborated to design high-priority observations and care actions adapted to the patient condition for each path. A nursing navigation content for each disease and treatment was created with nursing observations and care actions designed, including observations required by physicians. By appropriately matching each phase of the nursing navigation contents with each unit of the CPC, 778 digital path contents were developed for each disease and treatment. Since the nursing observation items required by the physician are defined as priority items, patient conditions that should be shared between the physician and nurse can be formulated efficiently and without omission in the nursing plan. (Figure 2)

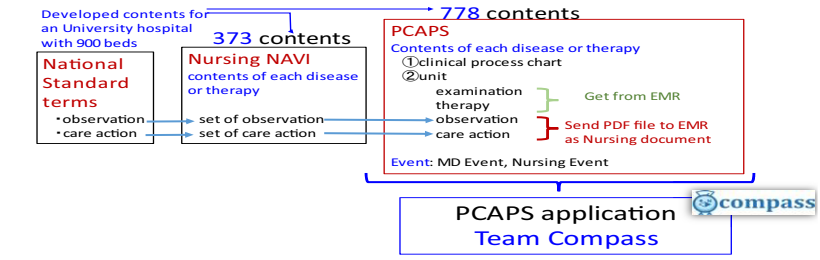


Figure 2. development of PCAPS contents

Step 3 (Development of the application system and operation design)
A clinical support system (Team Compass) with 778 different process path PCAPS has been developed. The main functions of the developed application system are to apply path contents to hospitalized patients, create nursing plans, input nursing observations and nursing actions performed, and refer to patient progress.

Physicians and surgeons select and order the appropriate path for their hospitalized patients from the EMRS. Nurses use Team Compass to create a nursing plan for each patient by selecting the most appropriate nursing observations and nursing actions for that patient from a list of candidates in the path applied to the patient. Observations and procedures are performed and implementation entries are made according to that nursing plan. (Figures 3 and 4)

Step 4 (Implementation)
In May 2019, a clinical support system linked to the EMRS was implemented in a 900-bed university hospital.

Step 5 (Evaluation)
Overtime hours for recording were compared and confirmed between March 2020, the end of the fiscal year when the system was familiarized with its operation, and the same month of the previous year. Unnecessary narrative records and inadequate sequential input were suggested as impediments to reducing overtime.

Step 6 (Verification)

After providing education on the aforementioned disincentive factors at a general hospital with approximately 400 beds, the same clinical support system was implemented in October 2020. Overtime hours for the following months, November and December 2020, were compared to the same months of the previous year.

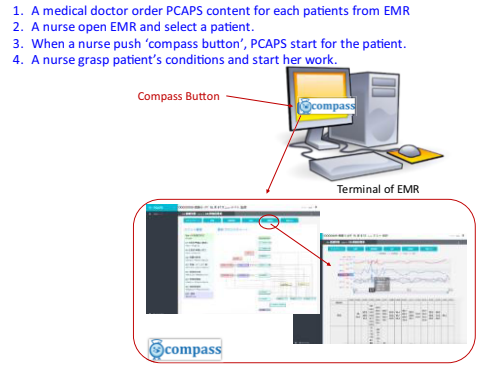


Figure 3. relationship between EMR and Application System (Team Compass)

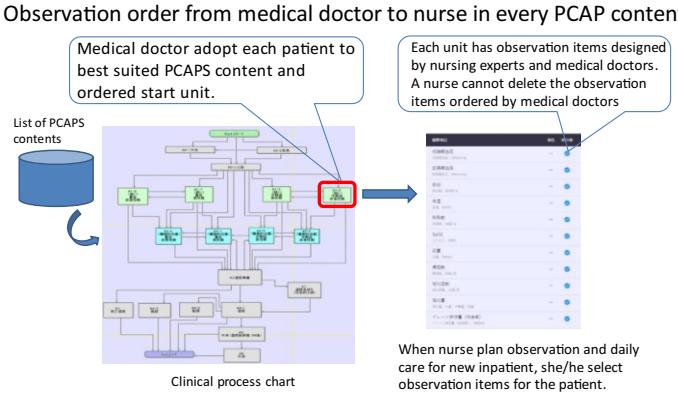


Figure 4. Application System (Team Compass)

3. Results

We used "overtime hours for nursing records" and "reduction rate" as indicators of the usefulness of this system.

The results were compared to the previous year's results in March, the end of the fiscal year at the 900-bed university hospital. These overtime hours were 2,944 hours 00 minutes in March 2019 and 2,141 hours 55 minutes in March 2020; a 27% reduction was indicated. The respective bed occupancy rates were 90.80% and 90.60%, with no difference; March 2018 and March 2019 are the time periods that can be said to be unaffected by COVID-19.

In a 400-bed general hospital, the months of November and December, one month after implementation, were compared to the previous year: 386 hours in November 2019

and 204.5 hours in November 2020, showing a 47% reduction; 366 hours in December 2019 and 214.5 hours in December 2020, showing a 41% reduction November/December 2019 and November/December 2020 are the periods affected by COVID-19.

4. Discussion

It was suggested that the application of 778 process paths, typified by disease and treatment, to all inpatients could reduce nurses' overtime hours for recording: 27% at a 900-bed university hospital and 40-50% at a 400-bed general hospital where disincentives were removed. In this system, clinically important observations are always performed. The creation of a system that allows physicians and nurses to share important patient conditions will contribute to improving the safety and quality of medical care. The 778 digital contents developed proved to be shareable across different hospitals. In the future, content management teams with rich clinical experience will be able to improve the content and share it on the cloud to achieve an efficient PDCA improvement cycle. It is expected that a knowledge ecosystem will be created by forming a community of hospital groups that have implemented this system

Acknowledgements

We are deeply grateful for the collaboration of the project members and organizations. This work was supported by JSPS KAKENHI (Grant-in-Aid for Scientific Research) No. 17H01608.

References

[1] Ministry of Health, Labor and Welfare: Toward the Realization of "Work Style Reform, https://www.mhlw.go.jp/stf/seisakunitsuite/bunya/0000148322.html, 2020
[2] Japan Nurses Association: Promotion of Work Style Reform for Nurses, https://www.nurse.or.jp/nursing/shuroanzen/hatarakikata/index.html, 2020
[3] Japan Federation of Healthcare Workers Unions: 2017 Report on the Results of the Survey on the Actual Working Conditions of Nursing Staff, Working Environment, 2017
[4] Satoko Tsuru, et.al.: Nursing Record Innovations Aimed at Harmonizing Structured Clinical Knowledge among Doctors and Nurses, “Digital Personalized Health and Medicine ", Studies in Health Technology and Informatics, 270, 2020, ISBN print: 978-1-64368-082-8, ISBN online: 978-1-64368-083-5., IOS Press, 5pages, 2020
[5] Satoko Tsuru, et.al.: a: Structuring Clinical Nursing Knowledge using PCAPS: Patient Condition Adaptive Path System, Proceedings of NI2009(Connecting Health and Humans IOS-Press ISBN 978-1-60750-024-7), scientific paper (2009), 391-395
[6] Satoko TSURU, et.al.: Reduction of Overwork Time of Nurses by Innovation of Nursing Records using Structured Clinical Knowledge, MEDINFO 2019: Health and Wellbeing e-Networks for All L. Ohno-Machado and B. Séroussi (Eds.) © 2019 International Medical Informatics Association (IMIA) and IOS Press, 1061-1064, 2019.
[7] Satoko Tsuru, Chitose Watanabe - the supervision and compilation: Clinical nursing in IT era -Navigator of thinking process for nursing-, Japan Standards Association (Tokyo), 2011 (Japanese)
[8] Satoko Tsuru: Recent concrete approach for medical security - Structuring of the Clinical Knowledge for the Patient Safety using PCAPS-, Quality, 39(4) 60-67, 2009 (Japanese)
[9] Paul R Ward, et.al.: ‘Waiting for’ and ‘waiting in’ public and private hospitals: a qualitative study of patient trust in South Australia, BMC Health Serv Res, 2017 May 5;17(1):333

Challenges of Trustable AI and Added-Value on Health
B. Séroussi et al. (Eds.)

doi:10.3233/SHTI220515

Healthcare Professionals' Experiences of the Work Environment After Patients' Access to Their Electronic Health Records - A Qualitative Study in Primary Care

Irene MULI[a,1], Isabella SCANDURRA [b], Åsa CAJANDER [c], and Maria HÄGGLUND[a]
[a] *Healthcare Sciences and E-health, Dept of Women's and Children's Health, Uppsala University, Sweden*
[b] *Informatics, School of Business Örebro University, Örebro, Sweden*
[c] *Department of Information Technology, Uppsala University, Sweden*

Abstract. For healthcare personnel, the work environment is already challenging, and when eHealth systems are introduced they are often considered to further add to the complexity. This paper investigates the impact of patients' access to their electronic health records on healthcare professionals' work environment in a primary care setting in Sweden. A work environment theory-driven analysis, focusing on perceived demands, control and support, was conducted on 14 semi-structured interviews of different primary care professionals. The professionals expressed a slight increase in demands, loss of control and some increase and decrease of support. This study discusses insights on how patients' access to health records can have an impact on healthcare professionals' work environment.

Keywords. Patient Accessible Electronic Health Records, Work Environment, Primary Care, Demand-Control-Support model

1. Introduction

Patient Accessible Electronic Health Records (PAEHRs) have been implemented in many countries [1–3]. For healthcare personnel, the work environment is already challenging [4], and more eHealth systems can further add to the complexity of the healthcare work environment [5]. Previous research has looked into healthcare staff and the PAEHR's effects on their work environment [6–8]. However, no studies have used a work environment theory-driven analysis [9] and few studies focused on primary care.

In the 1970's Robert Karasek developed a model for analyzing work-related stressors associated with cardiovascular illness. His 'demand and control model' was thereafter further developed together with Töres Theorell and is now one of the most widely used models for explaining psycho-social work conditions and their effects on health [9]. This model suggests that the combination of perceived *demands* and perceived *control* at work is a determining factor for stress. High job strain, i.e., high demands in combination with low decision latitude and low social *support,* are associated with the highest risks for health problems [10]. High demands usually are not a problem if

[1] Corresponding Author, Irene Muli, Uppsala universitet, Kvinnor och barns hälsa, Akademiska Sjukhuset, 751 85 Uppsala, Sweden; E-mail: irene.muli@kbh.uu.se.

combined with high self-control over work situations, tools, and substantial social support from management and colleagues. A skilled worker can experience this as a challenging situation. They have complete control over the work conditions and planning and receives full support when needed. The work is efficient and sustainable. On the other hand, if high demands are not met by substantial control and social support, the situation will soon be dangerous. [11]

Emanating from the Demand-Control-Support theory [9], this paper investigates the impact of patients' access to their health records on healthcare professionals' work environment in a primary care setting in Sweden.

2. Methods

In this study 14 healthcare professionals were interviewed representing seven different professions (counselor [C1], dietitians [D1-2], nurses [N1-2], physicians [P1-3], physiotherapists [PT1-2], occupational therapists [OT1], and medical secretaries [MS1-3]) working at a primary care center in Region Stockholm, Sweden. The semi-structured interviews were performed six months after patients were given access to the PAEHR through the national patient portal 1177.se [12]. Study participants were recruited during workplace meetings and informed consent was retrieved before the interviews. Interviews were performed face-to-face at the primary care center, recorded and lasted 30-45 min. The interview questions were related to the professionals' work in relation to the PAEHR. The interviews were transcribed and analyzed. The study has ethical approval from the Ethical Review Board in Stockholm (2017/1028-31).

The analysis was conducted with focus on effects of the implementation. Information coded under 'effects' was thereafter coded into three sub codes, *demand, control and support* based on the theoretical reference framework [9] which aims to provide an in-depth understanding of professionals' work environment. The other main codes were related to factors influencing implementation, the implementation processes, thoughts about suggested functions and about related functions.

3. Results

There was a general view among the respondents that little has changed due to and after the implementation of PAEHR. The patients seemed to still want their healthcare records printed on paper, although there was a hope PAEHR implementation would reduce the printing workload. Some respondents reported that they were trying to encourage patients to go online and look at their own PAEHR. There was also a sense that it was hard to reach out to patients about reading their PAEHR.

3.1. Demand - changes in how the PAEHR is documented

Despite prior requirements, some respondents were more aware of writing the notes on time and more comprehensively after the launch of PAEHR, whereas some reported no effects on their writing style. Being more considerate so the notes are comprehensible to more patients was described as initially being mentally demanding and time-consuming. There was also a view that this linguistic consideration was only in the beginning and that the professionals had turned back to how they used to write. Respondents

experienced pressure to write notes within the time-frame, which was described as challenging due to already existing time constraints.

The respondents reported increased documentation, e.g. adding more details to notes and writing down all interactions to avoid patients complaining of missing information.

''When writing perhaps you think a bit more before you express yourself and write. I try to be clear because I think that someone might read this. Maybe more now than before'' [D2]

3.2. Demand - increased workload due to patients' reactions to PAEHR

There was an experience of a slight increase in phone calls from patients regarding what is written in the records. Patients had e.g. questioned or refuted their diagnoses, how the notes were phrased by the healthcare professionals, notes written at the center and in one case notes from another center. There had also been a request from a patient to change old notes. Patients had also questions about their lab results, due to lack of understanding parts of it, or wondering how to move forward. Some of the patients' questions were experienced as too difficult to answer or irrelevant. There were also two cases where patients had booked a consultation only to get an explanation after reading their lab results, or to have their records read by a healthcare professional since they had not managed to read them online themselves.

''My experience is that the patients are very demanding, they might read some lab results online and wonder what they should do. A lot of complicated questions that we can't answer. It can be hard. We encounter a lot since we are the first contact'' [N2]

3.3. Demand - increased discomfort in interaction with patients and their informal carers

One respondent expressed feeling some uneasiness after unintentionally revealing a serious diagnosis to a patient that the patient was not aware of, assuming the patient had already read about it online. In another example a respondent felt some discomfort when a new patient from another primary care center after receiving lab results had requested a telephone consultation instead of a physical consultation. There has also been an instance where a relative had read a patient's health record which evoked discomfort for the interviewee.

''I did not know [them], I had never met [them], [they] had another physician, it did not feel right and I was supposed to give an evaluation on the phone, it did not feel right'' [P2]

3.4. Control - reduced control of planning and organization of work

The pressure to complete the documentation quickly after an appointment can also be perceived as a loss of control over when or how fast the notes are written. Reduced control was also experienced as they feared that the record lost some of its purpose to communicate medical information to other healthcare professionals, when writing the notes in a way that the patient can understand, and since the notes might be viewed by patients.

''This and that should be done, and it does not really work that way, we work with humans and humans get sick and there are no substitutes and the workload is different at different times'' [PT 2]

A potential reduced control over patient supervision was raised where there had been a case where patients after reading their test-results online canceled routine check-ups without consulting with the healthcare professionals. This was considered a risk as other examinations were planned for the check-up. Some respondents also felt an insecurity on how patients were informed, since unsigned information could be made available to patients when it comes to serious diagnoses. Another respondent had to call patients with suspicion of a serious diagnosis immediately, since they were aware the patient could see this.

3.5. Support - reduced support to coworkers and increased support in interaction with patients

The implementation of the PAEHR was said to work as a reminder to write the records on time, more precisely and correctly, but there was also a worry that the changed way of writing makes the note less understandable to other professionals and may thus be less of a support to them. Some respondents also stated that they could not to communicate freely through the records since patients could read them.

One respondent mentioned using the documentation to improve communication with patients, by documenting all conversations and referring patients to them when they try to refute previous agreements. Another also reported recommending the patients to access the PAEHR when they requested their healthcare records on paper.

''Sometimes when they request copies of their records you can say that you can go in and look at them yourself.'' [MS3]

4. Discussion and Conclusion

Although there was a general sense that little had changed due to the implementation of PAEHR, the primary healthcare professionals did report some increase in demand and slight increase of support but also some loss of control and support as previously reported. [13, 14]

According to the demand-control-support model, the reports of increased demand, loss of control and support would indicate an increased risk for health problems among the healthcare professionals. This since if the worker does not have control of work conditions and planning, does not have usable tools and feels exposed if things go wrong, the work will be unhealthy [11]. Such work situations are associated with high stress. In this extreme, health risks of different nature are common and people do not withstand the situation for long [11]. However, what is truly an impact of the introduction of the PAEHR in this case is a bit unclear since some of the reports on demand on documentation is said not to be new and that patients also prior to the implementation were able to request their healthcare records on paper. All records are now more accessible, hence the professionals experience a loss of control over the records.

There have been reports of the PAEHR supporting the communication with patients but also an increase of demand from patients. This increase in involvement by the patient is not necessarily negative, rather active patients have better health outcomes [15]. Patient involvement can also be viewed as a control function, where the patients could support the professionals by assuring that the information on the records is correct, but rather there is almost a sense of fear among healthcare professionals. Further studies are needed to investigate the professionals' perspective on the PAEHRs' potential to

improve the work environment. Reflexivity and other potential bias will be discussed in future longer publications.

In conclusion, e-health solutions that improve patient empowerment and patients' access to information are important, but when implemented we also need to consider their impact on healthcare professionals' work environment. When introducing new e-health solutions assessments using the demand-control-support model could be carried out to identify and reduce stress factors. Finding innovative ways to empower patients while at the same time not increasing demands or reducing support for healthcare professionals will be essential in the future.

Acknowledgement

This work is partly funded by FORTE – the Swedish Research Council for Health, Working Life and Welfare through the research project "PACESS" (2016-00623). AFA Insurance also provided funding for the analysis through the research project "ePrIm" (190210).

References

1. Essén A, Scandurra I, Gerrits R, Humphrey G, Johansen MA, Kierkegaard P, Koskinen J, Liaw S-T, Odeh S, Ross P. Patient access to electronic health records: differences across ten countries. Health Policy Technol. 2018;7(1):44–56.
2. Delbanco T, Walker J, Bell SK, Darer JD, Elmore JG, Farag N, Feldman HJ, Mejilla R, Ngo L, Ralston JD. Inviting patients to read their doctors' notes: a quasi-experimental study and a look ahead. Ann Intern Med. 2012;157(7):461–70.
3. Moll J, Rexhepi H, Cajander \AAsa, Grünloh C, Huvila I, Hägglund M, Myreteg G, Scandurra I, \AAhlfeldt R-M. Patients' Experiences of Accessing Their Electronic Health Records: National Patient Survey in Sweden. JMIR. 2018 Nov;20(11):e278.
4. Fiabane E, Giorgi I, Sguazzin C, Argentero P. Work engagement and occupational stress in nurses and other healthcare workers: the role of organisational and personal factors. J Clin Nurs. 2013;22(17–18):2614–24.
5. Cajander Å, Larusdottir M, Hedström G. The effects of automation of a patient-centric service in primary care on the work engagement and exhaustion of nurses. Qual User Exp. 2020;5(1):1–13.
6. Åkerstedt US, Cajander \AAsa, Moll J, \AAlander T. On threats and violence for staff and patient accessible electronic health records. Cogent Psychol. 2018;5(1):1518967.
7. Cajander Å, Moll J, Englund S, Hansman A. Medical Records Online for Patients and Effects on the Work Environment of Nurses. In: MIE. 2018. p. 271–5.
8. Moll J, Cajander Å. Oncology health-care professionals' perceived effects of patient accessible electronic health records 6 years after launch: A survey study at a major university hospital in Sweden. Health Informatics J. 2019;1460458219881007.
9. Karasek R, Theorell T. Healthy Work: Stress Productivity And The Reconstruction Of Working Life. Basic Books; Reprint edition (April 12, 1992); 1992.
10. Johnson JV, Hall EM. Job strain, work place social support, and cardiovascular disease: a cross-sectional study of a random sample of the Swedish working population. Am J Public Health. 1988;78(10):1336–42.
11. Wallgren LG, Hanse JJ. The impact of job characteristics and motivators on perceived stress among information technology (IT) consultants. Ergon Open J. 2010;3(1).
12. Hägglund M, Scandurra I. Patients' Online Access to Electronic Health Records: Current Status and Experiences from the Implementation in Sweden. In: MedInfo. 2017. p. 723–7.
13. Tapuria A, Porat T, Kalra D, Dsouza G, Xiaohui S, Curcin V. Impact of patient access to their electronic health record: systematic review. Inform Health Soc Care. 2021 Jun 2;46(2):194–206.
14. Turner A, Morris R, McDonagh L, Hamilton FL, Blake S, Farr M, et al. Unintended consequences of patient online access to health records: a qualitative study in UK primary care [Internet]. 2022. Available from: https://www.researchsquare.com/article/rs-1256673/v1 ry care''
15. Hibbard JH, Greene J, Overton V. Patients with lower activation associated with higher costs; delivery systems should know their patients' "scores." Health Aff Proj Hope. 2013 Feb;32(2):216–22.

Challenges of Trustable AI and Added-Value on Health
B. Séroussi et al. (Eds.)

doi:10.3233/SHTI220516

Are Personal Health Records (PHRs) Facilitating Patient Safety? A Scoping Review

Amanda L. JOSEPH[a,1], Helen MONKMAN[a], Andre W. KUSHNIRUK[a] and Elizabeth M. BORYCKI[a,b]
[a]*School of Health Information Science, University of Victoria, Canada*
[b]*Michael Smith Foundation for Health Research, Canada*

Abstract. Personal Health Records (PHRs) are poised to improve patient safety, however the mechanism(s) in which they improve safety is not clear. To this end, we conducted a scoping review with the following objectives: 1) explore the extent of the evidence that PHRs improve patient safety, 2) determine where PHR research has been done per International Medical Informatics Association (IMIA) Represented Region [1], 3) to identify the PHR naming convention(s) used per IMIA Region [1]. The findings revealed that there is limited evidence that PHRs improve patient safety. The results also revealed heterogeneity in PHR nomenclature and how they were used in healthcare settings. However, the overarching theme of the study, was that future research is needed to ensure that PHRs are designed and used in a patient safety context with human factors and usability considerations.

Keywords. Personal health record, patient safety, patient portal, health informatics

1. Introduction

Personal Health Records (PHRs) have gained market prominence as healthcare technologies, poised to improve patient safety. PHRs are electronic applications that enable authorized citizens to access, share, and manage health information in secure and confidential environments [2]. To understand their true value, PHRs must be contextualized as dynamic heterogeneous tools. Currently PHRs lack nomenclative consistency, functionality and design standardization. They have range from web-based applications, stand-alone systems, to programs tethered to a patient electronic health record (EHR) or electronic medical record (EMR) [2]. Consequently, limited longitudinal and empirical data is available to assess PHRs from broad perspectives and thus evidence of their impact on patient safety has remained unclear. The objectives of this scoping review were to: 1) establish the extent of the evidence that PHRs improved patient safety, 2) determine where PHR research has been done per International Medical

[1] Corresponding author, Amanda L. Joseph University of Victoria, British Columbia, Canada.
E-mail: amandalynnjoseph@uvic.ca

Informatics Association (IMIA) Represented Region [1], 3) to identify the PHR naming convention(s) used per IMIA Region [2].

2. Methods

We conducted a scoping review following the Arksey and O'Malley framework [3] using the PubMed® database, with the search terms "Personal Health Record" and "Patient Safety." PubMed® was utilized as it provided a robust sample of currently available published literature. To support interpretive consistency, prior to screening the articles, the researchers defined the terms PHR and patient safety. For the purposes of this study, we defined PHRs as the variety of electronic applications that enabled authorized citizens to access, share and manage health information in secure and confidential environments [2]. Thus, various terms and technologies that enabled citizens to modify and access their health information were categorized under the PHR umbrella. We defined patient safety as the opportunity to detect errors or risks in the healthcare system and remedy them to provide optimal patient outcomes [4]. Following this, two researchers conducted a first screen of the search results by title and abstract in Covidence®, then applied the following inclusion criteria: English articles (with abstracts) published between 2011-2021; PHR articles with a relationship to patient safety; PHRs tethered to EHRs or EMRs; PHRs with a different naming convention in which the data was maintained or modifiable by the patient or authorized delegate; electronic PHRs; PHR prototypes tested with patient populations. Articles that did not satisfy the inclusion criteria were excluded from the study. Once the final article sample was established, the authors completed a full text review of all remaining articles and tabulated the findings in a data extraction table (Table 1) for thematic analysis. The articles were iteratively assessed by two researchers based on the inclusion criteria and conflicts were resolved through discussion and consensus. A perceived limitation of the study, was the inclusion of electronically available English articles only and therefore other relevant articles may have been omitted based on this criteria. Additionally, as the intent of a scoping review is to provide a holistic assessment of the current state of the literature, all articles that satisfied the inclusion criteria were included, regardless of study design and methodological quality [3,5]. An ethics review was not required, as this was an assessment of publicly available literature.

3. Results

The search yielded 402 articles, with no duplicates and thus, all 402 articles were included in the first screen and assessed by title and abstract. From this, 336 articles were excluded as they did not meet the inclusion criteria and a full text review of the remaining 66 articles was done. After the 66 articles were read in full, another 44 articles were excluded and resulted in a final inclusion of 22 articles[2]. Of those 22 articles, 13 showed evidence that PHRs facilitated patient safety. We extracted the data (Table 1) using the following categories: PHR naming convention, IMIA Region [1], description of safety impact, thematic category. The four thematic categories identified in the literature were:

[2] For a complete reference list of all included articles, please contact the corresponding author.

1) shared decision making, 2) communication challenges, 3) medication safety, 4) usability and design challenges. Lastly, to facilitate comparative analysis, study specific acronyms were created for the PHR naming conventions, thematic categories and the IMIA Region [1] acronyms were utilized (Table 1).

Table 1. Data extraction

Naming convention	IMIA Region [1]	Description of safety impact	Thematic category	Reference
PHR, PP, EPR	APAMI, EFMI, MENAHI, NAMI	Health information systems have helped citizens improve their health status and manage their care with health professionals.	SDM	[6]
PHR	NAMI	PHR improved medication reconciliation self-management with patients and providers.	MS	[7]
PWP	NAMI	Portal communication was safe and effective, but unread message notifications posed risks.	CC	[8]
PP	N/A	Patient portals enabled discovery of medical errors and improved medication adherence.	MS	[9]
EPP	EFMI	42% of participants identified inaccuracies in clinical documents.	SDM	[10]
PP	EFMI	Patients had difficulties interpreting and understanding laboratory test results.	CC	[11]
PP	NAMI*	Patients and families identified errors and recognized quality problems in their records.	SDM	[12]
ACP	NAMI	Participants found medical terms and acronyms in clinical notes hard to understand.	CC	[13]
IP	NAMI	The portal facilitated recognition of medication errors.	MS	[14]
PP	NAMI	Participants reported errors online and 93% found that reporting mistakes improved safety.	MS	[15]
OPHR	NAMI	The PHR tool improved safety and decreased patient-reported medication errors.	MS	[16]
PMR	N/A	Patients were able to correct mistakes and discrepancies in their medical file themselves.	SDM	[17]
PHR	NAMI	Interception of medication dosing errors was achieved by patients accessing their data.	MS	[18]

*Assumption: authors are affiliated with NAMI Region institutions.

Naming convention(s):Personal health record (PHR), patient portal (PP), electronic patient record (EPR), patient web portal (PWP), electronic patient portal (EPP), acute care portal (ACP), inpatient portal (IP), online PHR (OPHR), personal medical record (PMR). Region: Asia Pacific (APAMI), European (EFMI), Middle East and North African (MENAHIA), North America (NAMI). Thematic categories: Shared decision making (SDM), communication challenges (CC), medication safety (MS), usability and design challenges (UDC).

The results of the thematic analysis indicated that 9 of the 22 of the articles represented medication safety issues and how PHR interventions impacted medication administration (e.g., reconciliation, error detection, dosage). Whereas 6 articles described communication challenges between providers and citizens with PHR use. Shared decision making in which citizen and provider collaboration was established through frequent communication and open medical record access, was present in 5 articles. Lastly, 2 articles described usability and design challenges which prevented citizens from using PHRs safely, effectively and enjoyably [19]. The findings (Table 1) revealed that heterogeneity existed in: PHR naming convention(s), where PHR research has been done, how PHRs impacted safety, thematic category. There was evidence that PHRs facilitated patient safety in 13 articles, however, 9 articles failed to provide clear evidence that PHR use improved patient safety (Figure 1).

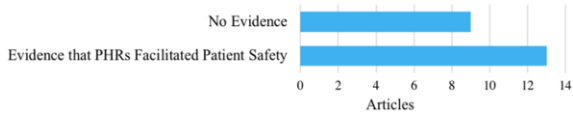


Figure 1. Evidence of patient safety in the literature

The findings revealed that there is much to learn about the relationship between PHRs and patient safety. Additionally, it was discovered that future work is required to establish safety parameters in PHR design and evaluation. The PHR nomenclature varied (Figure 2) and with 5 articles, the patient portal was the most prominent naming convention. Additionally, the IMIA Region [1] representation was diverse (Figure 3), 9 articles described the NAMI region as the most prominent region for PHR research.

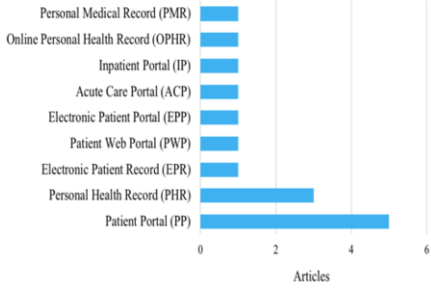


Figure 2. Varied naming conventions

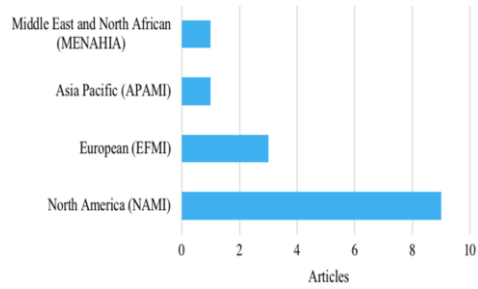


Figure 3. PHR research per IMIA Region [2]

4. Discussion and Conclusion

This scoping review presented a range of PHR naming conventions, research representation by IMIA Region [1]. As evidenced by the findings, the relationship between PHRs and patient safety has only been partially established. Therefore, the integration of usability, human factors and human cognitive processing abilities [20] into PHR design and concept planning, could result in safer systems. Moreover, utilizing patient journey mapping activities [21] in PHR design and implementation planning could improve patient outcomes. Such activities could illustrate the varied clinically related intersections that exist between citizens and providers that could compromise patient safety. A broader analysis is required to further explore the current state of PHRs in the global marketplace. There is a need for a generalizable framework to evaluate PHRs from stratified perspectives (e.g., micro, meso, macro, multi [22]) to improve design, safety, functionality and adoption.

Acknowledgement Amanda L. Joseph has received funding from the Natural Science and Engineering Research Council of Canada (NSERC) Visual and Automated Disease Analytics (VADA) Program.

References

[1] IMIA Represented Regions [Internet]. IMIA. [cited 2022 Jan 19]. Available from: https://imia-medinfo.org/wp/regional-members/.

[2] Chumbler NR, Haggstrom D, Saleem JJ. Implementation of health information technology in Veterans Health Administration to support transformational change: telehealth and personal health records. Med Care. 2011 Dec;49 Suppl:S36-42.
[3] Arksey H, O'Malley L. Scoping studies: towards a methodological framework. Int J Soc Res Methodol.2005;8(1):19-32.
[4] General Patient Safety [Internet]. Canadian Patient Safety Institute.[cited 2022 Jan 19]. Available from: https://www.patientsafetyinstitute.ca/en/Topic/Pages/General-Patient-Safety.aspx.
[5] Tricco AC, Lillie E, Zarin W, O'Brien KK, Colquhoun H, Levac D, Moher D, Peters MDJ, Horsley T, Weeks L, Hempel S, Akl EA, Chang C, McGowan J, Stewart L, Hartling L, Aldcroft A, Wilson MG, Garritty C, Lewin S, Godfrey CM, Macdonald MT, Langlois EV, Soares-Weiser K, Moriarty J, Clifford T, Tunçalp Ö, Straus SE. PRISMA Extension for Scoping Reviews (PRISMA-ScR): Checklist and Explanation. Ann Intern Med. 2018 Oct 2;169(7):467-473.
[6] Borycki EM, Househ MS, Kushniruk AW, Nohr C, Takeda H. Empowering Patients: Making Health Information and Systems Safer for Patients and the Public. Contribution of the IMIA Health Informatics for Patient Safety Working Group. Yearb Med Inform. 2012;7:56-64.
[7] Chrischilles EA, Hourcade JP, Doucette W, Eichmann D, Gryzlak B, Lorentzen R, Wright K, Letuchy E, Mueller M, Farris K, Levy B. Personal health records: a randomized trial of effects on elder medication safety. J Am Med Inform Assoc. 2014 Jul-Aug;21(4):679-86.
[8] Crotty BH, Mostaghimi A, O'Brien J, Bajracharya A, Safran C, Landon BE. Prevalence and Risk Profile Of Unread Messages To Patients In A Patient Web Portal. Appl Clin Inform. 2015 Jun 12;6(2):375-82.
[9] Dendere R, Slade C, Burton-Jones A, Sullivan C, Staib A, Janda M. Patient Portals Facilitating Engagement With Inpatient Electronic Medical Records: A Systematic Review. J Med Internet Res. 2019 Apr 11;21(4):e12779.
[10] Fitzsimons M, Power K, McCrea Z, Kiersey R, White M, Dunleavy B, O'Donoghue S, Lambert V, Delanty N, Doherty CP. Democratizing epilepsy care: Utility and usability of an electronic patient portal. Epilepsy Behav. 2021 Sep;122:108197.
[11] Fraccaro P, Vigo M, Balatsoukas P, van der Veer SN, Hassan L, Williams R, Wood G, Sinha S, Buchan I, Peek N. Presentation of laboratory test results in patient portals: influence of interface design on risk interpretation and visual search behaviour. BMC Med Inform Decis Mak. 2018 Feb 12;18(1):11.
[12] Gerard M, Fossa A, Folcarelli PH, Walker J, Bell SK. What Patients Value About Reading Visit Notes: A Qualitative Inquiry of Patient Experiences With Their Health Information. J Med Internet Res. 2017 Jul 14;19(7):e237.
[13] Grossman LV, Creber RM, Restaino S, Vawdrey DK. Sharing Clinical Notes with Hospitalized Patients via an Acute Care Portal. AMIA Annu Symp Proc. 2018 Apr 16;2017:800-809.
[14] Kelly MM, Hoonakker PL, Dean SM. Using an inpatient portal to engage families in pediatric hospital care. J Am Med Inform Assoc. 2017 Jan;24(1):153-161.
[15] Lam BD, Bourgeois F, Dong ZJ, Bell SK. Speaking up about patient-perceived serious visit note errors: Patient and family experiences and recommendations. J Am Med Inform Assoc. 2021 Mar 18;28(4):685-694.
[16] Schnipper JL, Gandhi TK, Wald JS, Grant RW, Poon EG, Volk LA, Businger A, Williams DH, Siteman E, Buckel L, Middleton B. Effects of an online personal health record on medication accuracy and safety: a cluster-randomized trial. J Am Med Inform Assoc. 2012 Sep-Oct;19(5):728-34.
[17] Vermeir P, Degroote S, Vandijck D, Van Tiggelen H, Peleman R, Verhaeghe R, Mariman A, Vogelaers D. The patient perspective on the effects of medical record accessibility: a systematic review. Acta Clin Belg. 2017 Jun;72(3):186-194.
[18] Woollen J, Prey J, Wilcox L, Sackeim A, Restaino S, Raza ST, Bakken S, Feiner S, Hripcsak G, Vawdrey D. Patient Experiences Using an Inpatient Personal Health Record. Appl Clin Inform. 2016 Jun 1;7(2):446-60.
[19] Kushniruk AW, Patel VL. Cognitive and usability engineering methods for the evaluation of clinical information systems. Journal of Biomedical Informatics. 2004;37(1):56-76.
[20] Joseph AL, Borycki EM, Kushniruk AW. Alert fatigue and errors caused by technology: A scoping review and introduction to the flow of cognitive processing model. Knowledge management & e-learning. 2021;13(4):500-521.
[21] Joseph AL, Kushniruk AW, Borycki EM. Patient journey mapping: Current practices, challenges and future opportunities in healthcare. Knowledge management & e-learning. 2020;12(4):387-404.
[22] Joseph AL, Stringer E, Borycki EM, Kushniruk AW. Evaluative Frameworks and Models for Health Information Systems (HIS) and Health Information Technologies (HIT). Studies in health technology and informatics. 2022; 289: 280 - 285.

Challenges of Trustable AI and Added-Value on Health
B. Séroussi et al. (Eds.)

doi:10.3233/SHTI220517

Goupile: A New Paradigm for the Development and Implementation of Clinical Report Forms

Niels MARTIGNENE[a,b,1], Ali AMAD[c], Julie BELLET[d], Julien TABAREAU[d], Fabien D'HONDT[b,c], Thomas FOVET[b,c] and Antoine LAMER[a,e]
[a]*InterHop, F-59000 Lille, France*
[b]*Centre National de Ressources et de Résilience Lille-Paris (CN2R), Lille, France*
[c]*Univ. Lille, Inserm, CHU Lille, U1172 – LilNcog - Lille Neuroscience & Cognition, F-59000 Lille, France*
[d]*CHU de Lille, Pôle d'anesthésie-réanimation, F-59000 Lille, France*
[e]*Univ. Lille, CHU Lille, ULR 2694 - METRICS: Évaluation des Technologies de santé et des Pratiques médicales, F-59000 Lille, France*

Abstract. Despite the increasing computerization of hospital information systems, segments of patient care are still in paper format. Data extracted automatically from the hospital databases for one specific project are thus supplemented by data collected manually. Data collection tools are usually developed entirely, which requires computer knowledge and is tedious, or automatically from metadata or drag and drop controls, which is limiting in terms of functionality. To facilitate this manual collection, we developed a free and open-source tool for creating forms that does not require advanced computer skills, offers rich features, and is quickly implemented, tested and deployed. It was implemented for 15 projects and supported thousands of daily users for a complex interactive study at the national level.

Keywords. Clinical research; Electronic data capture; Electronic health record; Data collection

1. Introduction

Despite the increasing computerization of health facilities, health data are still being collected manually to supplement the computerized databases when conducted observational retrospective or prospective studies. There are three main types of tools for manually capturing and storing data: spreadsheets, Electronic Data Capture (EDC) systems and handmade applications.

Spreadsheets are part of office suite software. In the same manner as text editors or presentation programs, spreadsheets are prevalent because they are automatically installed on the workstation and easy to handle [1]. However, they can be modified by a

[1] Corresponding Author, Niels Martignène, CN2R, 103 Boulevard de la Liberté, F-59000, Lille, France; E-mail: niels.martignene@protonmail.com.

faulty manipulation, they do not support the generation of standardized and interoperable data, and it is particularly difficult to manually handle for large datasets with high number of rows and columns.

EDCs collect data in an electronic Clinical Report Form (eCRF) and are generally composed of a graphical user interface component for data entry, a validation component and a database [2]. There are two main ways to design a form with an EDC: a metadata-driven approach and a GUI approach. In the first case, a table with the list of required variables, types, labels and constraints are loaded into the EDC to automatically structure the visual interface and the database. In the second case, graphical widgets are manually positioned on the interface by drag and drop. The widgets are then configured to take into account the characteristics of each variable. Although they have been reported as a mean of improving research efficiency [3], both approaches often offer limited functionality, and proprietary license is a barrier for scientific projects without sufficient funding [4].

The last alternative consists in developing handmade applications from scratch. For this, a web technology (such as PHP, Node.js, etc.) is generally associated with a database management system, which control the user interface and data storage, respectively. This approach allows to develop forms with unlimited functionalities, but requires advanced programming skills and is time consuming.

The objective of this project is to provide a free and open-source tool facilitating the developer and end-user work during the eCRF design stage. The tool has to allow creating forms on the fly, immediately ready to be tested and used, with a minimum of code and coding skills.

2. Methods

We performed extensive interviews among our peers in medical, pharmaceutical and environmental sciences to assess their needs in term of data capture tools. To bridge the gap between spreadsheets, EDCs and handmade applications, our solution had to answer the following design goals: (i) enable the creation of the project, the design of the interface, and the formatting of the data structure and storage with a minimum of code; (ii) facilitate the development and interactions between developer and end-user, (iii) support standard widgets but also rich functionalities, (iv) view and export the data easily, (v) provide training, (vi) generate shareable forms, (vii) support an offline mode and mobile/tablet format, and (viii) be free and open-source.

To bridge the gap between existing approaches, we have developed a framework which supports the generation the interface, data structure, and storage in one line using a common programming language rather than a meta-language or a graphical library.

We have developed Goupile project using free and open-source technologies. The front-end is built using HTML/CSS/Javascript, while the back-end is implemented in C++. Data is stored on the server as JSON (JavaScript Object Notation) objects in a SQLite database. The choice of SQLite was made for three reasons: (i) we wanted to compile and use Goupile as an autonomous single-file binary, (ii) we use relational tables for other functionality, such as user management, and (iii) SQLite is a solid battle-tested database engine with an extensive test suite. Each record does not necessarily have the same variables, which allows the form to evolve and store new variables. Some data can be stored client-side with IndexedDB (offline mode).

The standard widgets are implemented with predefined Javascripts functions: number, text, dropdown list, radio button, checkbox, date. Other more specific widgets were developed according to user needs: choice (a set of horizontally arranged buttons), binary (choice with yes/no answer), likert scale, and computed variables. All widgets are configurable with optional arguments dealing with the mandatory of the field, the number of decimals, or the minimal and maximal values, for example. Optional arguments are passed to the function between curly braces (see Figure 1).

```
function ( "variable_name", "Label shown to user", {
    option1 : value,
    option2 : value
} )
```

Figure 1. Function template for the creation of a widget

3. Results

Goupile is composed of five main components:

- an administration panel to create projects, users and assign projects to users in defining permissions (code, entry, export, read what was read by other),
- an application editor to define the architecture of a project (the forms and the relationships between them) and its general configuration (date of start, style with a CSS sheet),
- a form editor to code the sections and widgets (label, type, options),
- an entry panel to display instantly the form and its visual widgets, which allows the user to enter data,
- a data overview module to display the data recorded and provide export function (in csv or xlsx formats), audit trail displays information on entry (user and date); a dashboard displays the number of records and number of complete cases.

The project is available in two formats, a ready-to-use package for a single form and a single user, and a complete solution for multiple forms and with different configuration options [5].

3.1. Development of the form

During the form design phase, the form editor may be used simultaneously with the entry panel: while coding the widget in the editor, the control is immediately displayed and usable right next to it (see Figure 2). At this stage, it is already possible to validate a record, test the logical structure of the form and visualize fictitious records.

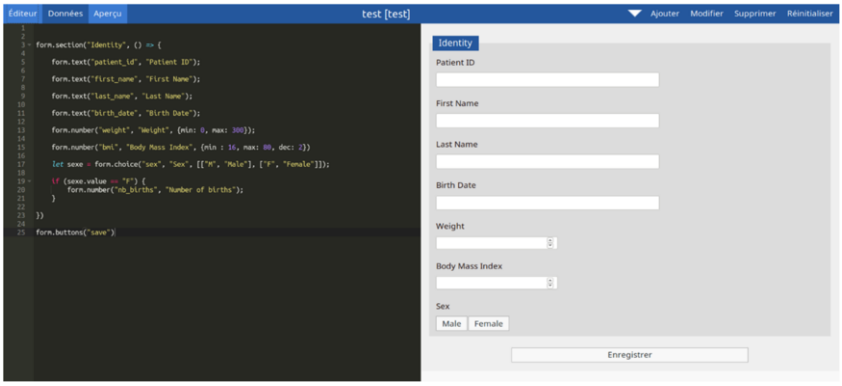


Figure 2. Form editor and entry panel. The left part of the interface is the form editor. The right part of the interface is the entry panel.

3.2. Daily use

During the data collection phase, the entry panel may be used alone, or with the data overview/export part (see Figure 3).

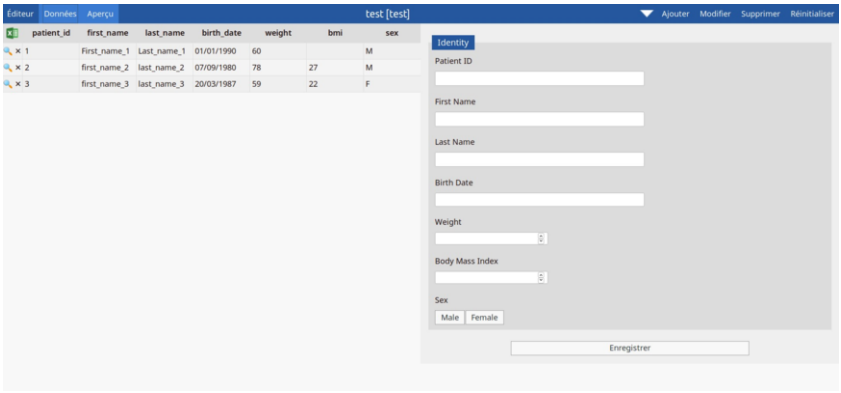


Figure 3. Data and form overview. The left part of the interface is the data overview. An export button allows exporting a CVS file. The right part of the interface is the entry panel.

3.3. Functionality

The main features are the support of 1-to-many relationships, an offline mode, responsive design for tablets and smartphones, error management, data tracking. Advanced users can also customize the form with conditions, loops and all that can be programmed with Javascript. Finally, the data is hosted on a secure server.

3.4. Projects

At this time, Goupile has been used for 15 projects, including 3 multi-center projects. The design of Goupile was flexible enough for us to implement the 17 modules of the MINI for DSM-V assessment online, despite the complex flow of the questions. We were also able to conduct a complex interactive study at the national level with Goupile, with

several neuropsychological assessments, where the users had to watch multiple videos and photos and answer to associated questions.

Several studies were developed on Goupile by non-programmers, who where able to do create most of the form each time. We typically intervened only at the end to fix minor mistakes and help to code the more complex conditions needed for some questions.

4. Discussion

We offer a tool that allows to quickly design and edit a form using ready-to-use Javascript functions, immediately visualize and test the form, and visualize/export collected data. The use of ready-made functions directly modifies the appearance of the form and the structure of the database, thus saving time compared to an application developed from scratch. These functions are developed in a common language (Javascript), rather than a meta-language or a GUI, which offers a number of customization possibilities for the eCRF. Finally, the hosting on a server, and the data quality rules are advantages compared to the spreadsheets. Goupile is provided free of charge, and the code is free and open-source (AGPL 3.0).

The tool was well received by the users, both for the form design and data collection phases. Further developments will address the support of direct mail, the use in clinical routine, and more powerful data monitoring capabilities.

5. Conclusions

Goupile presents a new paradigm for the quick development and implementation of an eCRF. Based on Javascript and ready-made functions, it allows users to easily design, edit, test and deploy an eCRF.

References

[1] Anderson NR, Lee ES, Brockenbrough JS, Minie ME, Fuller S, Brinkley J, Tarczy-Hornoch P. Issues in biomedical research data management and analysis: needs and barriers. J Am Med Inform Assoc. 2007 Jul-Aug;14(4):478-88. doi: 10.1197/jamia.M2114. Epub 2007 Apr 25. PMID: 17460139; PMCID: PMC2244904.

[2] Shah J, Rajgor D, Pradhan S, McCready M, Zaveri A, Pietrobon R. Electronic data capture for registries and clinical trials in orthopaedic surgery: open source versus commercial systems. Clin Orthop Relat Res. 2010 Oct;468(10):2664-71. doi: 10.1007/s11999-010-1469-3. PMID: 20635174; PMCID: PMC3049639.

[3] Rorie DA, Flynn RWV, Grieve K, Doney A, Mackenzie I, MacDonald TM, Rogers A. Electronic case report forms and electronic data capture within clinical trials and pharmacoepidemiology. Br J Clin Pharmacol. 2017 Sep;83(9):1880-1895. doi: 10.1111/bcp.13285. Epub 2017 Apr 22. PMID: 28276585; PMCID: PMC5555865.

[4] Open-Source Electronic Health Record Systems for Low-Resource Settings: Systematic Review. JMIR Med Inform. 2017 Nov 13;5(4):e44. doi: 10.2196/medinform.8131. PMID: 29133283; PMCID: PMC5703976.

[5] (n.d.). https://framagit.org/interhop/goupile (accessed January 20, 2022).

Challenges of Trustable AI and Added-Value on Health
B. Séroussi et al. (Eds.)

doi:10.3233/SHTI220518

Exploring the Digital Divide as a Barrier to Use of a Personal Health Record in the Elderly

Eliana FRUTOS[a,1], Juan DESCALZO[a], Giuliana COLUSSI[a], Romina RAPISARDA[a], Fernando PLAZZOTTA[a] and Daniel LUNA[a]
[a]*Health Informatics Department, Hospital Italiano de Buenos Aires, Argentina*

Abstract. The digital divide can hinder the ability of elderly patients to fully benefit from PHRs. They are "digital immigrants", not having the life-long exposure to technology as younger generations, as well as physical and cognitive disabilities. The aim of this study was to explore the digital divide as a barrier for the use of a PHR in older adults (> 69 years of age) and describe the use of a PHR in an elderly population in Argentina. We conducted a cross sectional study which included older adults who attended the Coronavirus vaccination campaign in 2021. Data were collected through a survey encompassing digital divide factors and use of the PHR. A total of 128 participants agreed to complete the survey, 60.15% reported using the PHR. We found a statistically significant correlation of education level, having a personal computer and internet access with PHR use. Concerning PHR users, 45.45% reported needing assistance to use it. Although the elderly population represents a large portion of patients, there is not enough research done on their use experience using eHealth solutions. There is pending work in the eHealth field to integrate these elders into current PHRs and help them enjoy their benefits.

Keywords. Personal Health Records, Digital divide, Elderly, Consumer Health Informatics.

1. Introduction

The digital divide is the gap between those who have access to and benefit from modern information technology and those who don't [1]. Originally, the focus was set to access, mainly to computers and an internet connection. Nowadays, this divide is understood as a gap in the ability to use and gain benefits from technology, not only limited by access [2]. There are many factors acting on this divide such as gender, socioeconomic level, education and age [3]. The elderly are a group particularly at risk, as they are "digital immigrants", not having the life-long exposure to technology as younger generations [4], as well as physical and cognitive disabilities [5].

The world's population is rapidly aging, primarily in developing countries. In 2019, there were 703 million people aged 65 years or older (9.1% of the global population), this number is projected to double to nearly 1.5 billion in 2050 [6]. The increasing proportion of older adults presses upwards pressure on overall healthcare spending, as

[1] Eliana Frutos, Health Informatics Department, Hospital Italiano de Buenos Aires, Tte. Gral. Juan D. Perón 4190, C1199ABB Buenos Aires, Argentina; E-mail: eliana.frutos@hospitalitaliano.org.ar.

the use of medical care services rises with age, as well as per capita costs of healthcare [7-8].

Personal health records (PHRs) may be a valuable tool for older adults to manage their personal health information and self-manage their health and conditions [9-10]. They have been shown to provide long-term positive net value for organizations [11]. Usings PHRs patients may view information like laboratory results and medication history, securely message their physicians, request prescription refills, and schedule appointments [12]. Factors influencing the digital divide in this population may hinder their possibilities to fully adopt PHRs and their benefits, this must be taken into account in design and implementation [5].

The goal of this study is to explore the digital divide as a barrier for the use of a PHR in older adults (> 69 years of age) and describe the use of a PHR in an elderly population in Argentina.

2. Methods

2.1 Setting

Our study took place at Hospital Italiano de Buenos Aires (HIBA), a community-based tertiary care hospital located in Buenos Aires, Argentina. HIBA is a HIMSS Level 7+ organization with an in-house developed health information system [13]. An integrated PHR has been available since 2007. Its main functionalities include: appointment scheduling, test results visualization, patient-practitioner secure messaging, health information sharing, medication management, and different teleconsultation modalities. At present, it has approximately 400,000 registered users. HIBA has its own health insurance called Plan de Salud (PS) with over 150,000 affiliates.

2.2 Design, participants and data collection

We conducted a cross sectional study which included older adults who attended the Coronavirus vaccination campaign between March and April 2021.

Eligible participants were included if they were older than 69 years old and affiliated to HIBA PS. Those physically or mentally unable to answer were excluded. Participants consented orally to participation.

Data were collected through a survey designed by an interdisciplinary team of health informatics specialists, sociologists and psychologists. The survey encompassed demographic characteristics, educational level, internet connection, access to electronic devices, use of PHR, need of assistance to use the PHR and functionality usage.

2.3 Ethical considerations

The research project was approved by the institutional ethics committee (CEPI # 4501). The study was performed in full agreement with current national and international ethical regulations.

3. Results

A total of 128 participants agreed to complete the survey. The median age was 77 (IQR 6) years old and the female sex was predominant (61.71%).

All surveyed had a HIBA PHR account but only 60.15% reported using it. Most non-users' PHRs were managed by someone else, usually a family member.

We explored variables related to the digital divide as internet access and electronic devices. We found a statistically significant correlation of having a personal computer ($p < 0.001$) and internet access ($p = 0.003$) with PHR use. Also evidenced with the variable education ($p = 0.03$).

Table 1. Baseline characteristics, digital divide and a comparison between users and non-users.

	Total = 128	Users (N=77)	Non-users (N=51)	p value
Baseline characteristics				
Female sex	61.71% (79)	68.83% (53)	50.98% (26)	0.52
Age, in years*	77.6 (IQR 6)	77 (IQR 6)	78 (IQR 6)	0.07
Education				0.03
Primary	14.84% (19)	11.68 % (9)	19.60% (10)	
Secondary	37.50% (48)	24.67% (19)	56.86% (29)	
Tertiary	23.44% (30)	28.57% (22)	15.68% (8)	
University or higher	24.22% (31)	35.06% (27)	7.84% (4)	
Digital divide				
Internet access	88.28% (113)	98.70% (76)	72.54% (37)	0.003
Electronic device access				
PC**	71.87% (92)	89.61% (69)	45.09% (23)	<0.001
Smartphone	92.18% (118)	97.40% (75)	84.31% (43)	0.22

*Median (Interquartile Range)
**PC=Personal computer

Concerning PHR users, 45.45% reported needing assistance to use it. Regarding PHR features, scheduling appointments (96.10%), review test results (84.41%) and teleconsultations (76.05%) were the most frequently used.

4. Discussion

In this study, we attempted to explore the digital divide in elderly users and non-users of a PHR and describe its use with the objective of rethinking effective ways to engage this population in the adoption of personal health records.

It is often assumed that the elderly population are non technology users [14], however our results showed that 88.28% had internet access, 71.87% had a personal computer and almost all the surveyed had smartphones. As the years go by this is a changing reality, the preconceptions about older adults from decades ago are no longer

true today. They have a higher level of connectivity, use smart devices and connect to the internet to keep in touch, socialize, manage their finances and consume media [15]. There is pending work in the eHealth field to integrate these elders into current PHRs and help them enjoy their benefits. This area is also critical when considering the possibilities of AI in PHRs, enhancing patient's experience (such as study interpretation support and presenting relevant information), understanding their perspective and attitudes towards AI and technology is key for successful implementation and AI-supported PHR enhancements.

When exploring the difference between PHR users and non-users, having a smartphone couldn't be correlated with PHR use, however having a personal computer was statistically significant ($p < 0.001$) and may be related with regular internet usage and a higher skill level regarding information technology. Computer literacy has been described as a common barrier for PHR use in works such as Jabour et al [16].

Another determinant for use is education level, our results ($p = 0.03$) are consistent with similar findings reported in the literature. Educational level is a known determinant of health. A higher education gives people the tools they need to access the health system and effective self-care. Advanced education may be a proxy for higher socioeconomic status. Both factors have been described as determinants of the digital divide [17].

Almost half of the participants who used the PHR (45.45%) said they needed some kind of assistance. While they may be aware of its functionalities and willing to use it they couldn't manage it by themselves completely. This division in assistance needed shows that this is an heterogeneous population regarding their usability and accessibility needs. It should be a strong point to take into consideration when designing PHRs: both accessibility options for the elderly and taking into account that not only the patients themselves may be users, but also families and caregivers. Similar findings have been reported by Kim et al [18].

Although the elderly population represents a large portion of patients, there is not enough research done on their use experience using eHealth solutions. Not fully understanding their needs and usage patterns may cause PHRs design to "not fit" properly into their daily lives, homes and needs. It is necessary to study the elderly in their different contexts. Many are living in nursing or retirement homes, which may be simpler to approach for researchers, in contrast with adults who live in their own homes. It's easy to group them into a single label, hiding their true complexity.

While there are many studies in the literature researching PHR use and the digital divide, most have a Northamerican or European setting. These could not adequately represent the population of developing countries such as Argentina, where the education and socioeconomic realities are different. Our research is relevant because it is the first to explore PHR use and the digital divide in an elderly population in our local context.

This study has some limitations. This is a single-center study and sample size ($n = 128$) was limited for practical reasons, which impacts the generalizability of this work. The older adult participants may not be representative of the general older adult population affiliated to PS, nor the general elderly population of Argentina, as private medicine patients represent a wealthy strata. On another note, patients had to be physically able to travel to the hospital to receive the vaccine, these could be sources of bias in our sample.

As a future line of work, we intend to explore more aspects of the digital divide, PHR user satisfaction and usability, expand our sample size and collaborate with other medical centers. This study is foundational to future work and research in the area.

5. Conclusion

The digital divide can hinder the ability of elderly patients to fully benefit from PHRs. The findings from this study revealed that many digital divide factors are related with PHR use among older adults, as has been reported in the literature. More research is necessary, applying a usability approach, to explore their needs and patterns of use, as well as their capabilities. In the future, more of our users will be older adults, it is necessary that health informatics professionals begin to take them into account as relevant eHealth consumers.

References

[1] Delfino G, Sosa F, Zubieta E. (2017). Internet usage in Argentina: Gender and age as variables associated to the digital divide. Investigación y Desarrollo, 25(2), 100-123.
[2] Van Dijk JA. (2005). The deepening divide: Inequality in the information society. Sage Publications.
[3] Guillén MF, Suárez S. L. (2005). Explaining the global digital divide: Economic, political and sociological drivers of cross-national Internet use. Social forces, 84(2), 681-708.
[4] Vidal E. (2019, October). Digital Literacy Program: Reducing the Digital Gap of the Elderly: Experiences and Lessons Learned. In 2019 International Conference on Inclusive Technologies and Education (CONTIE) (pp. 117-1173). IEEE.
[5] Lober WB, Zierler B, Herbaugh A, Shinstrom SE, Stolyar A, Kim EH, et al. (2006). Barriers to the use of a personal health record by an elderly population. In AMIA Annual Symposium Proceedings (Vol. 2006, p. 514). American Medical Informatics Association.
[6] United Nations, Department of Economic and Social Affairs, Population Division (2019). World Population Ageing 2019: Highlights (ST/ESA/SER.A/430)
[7] World Health Organization. (2015). World report on ageing and health. World Health Organization. https://apps.who.int/iris/handle/10665/186463
[8] Pilger C, Menon MU, Mathias TA. D. F. (2013). Health services use among elderly people living in the community. Revista da Escola de Enfermagem da USP, 47, 213-220.
[9] Kruse CS, Argueta DA, Lopez L, Nair A. (2015). Patient and provider attitudes toward the use of patient portals for the management of chronic disease: a systematic review. Journal of medical Internet research, 17(2), e3703.
[10] Tavares J, Oliveira T. New Integrated Model Approach to Understand the Factors That Drive Electronic Health Record Portal Adoption: Cross-Sectional National Survey. J Med Internet Res. 2018;20: e11032.
[11] Kaelber D, Pan EC. (2008). The value of personal health record (PHR) systems. In AMIA Annual Symposium Proceedings (Vol. 2008, p. 343). American Medical Informatics Association.
[12] Fraccaro P, Vigo M, Balatsoukas P, Buchan IE, Peek N, van der Veer SN. (2018). The influence of patient portals on users' decision making is insufficiently investigated: a systematic methodological review. International journal of medical informatics, 111, 100-111.
[13] Plazzotta F, Luna D. (2015). Health information systems: integrating clinical data in different scenarios and users. Revista peruana de medicina experimental y salud publica, 32(2), 343-351.
[14] Nahm ES, Son H, Yoon JM. (2020). Older adults' use of patient portals: Experiences, challenges, and suggestions shared through discussion board forums. Geriatric Nursing, 41(4), 387-393.
[15] Pew Research Center. Technology Use Among Seniors. 2017. http://www. pewinternet.org/2017/05/17/technology-use-among-seniors/. Accessed March 30, 2018.
[16] Jabour A, Jones JF. (2013, July). Facilitators and barriers to patients' engagements with personal health records: systematic review. In International Conference on Universal Access in Human-Computer Interaction (pp. 472-481). Springer, Berlin, Heidelberg.
[17] Nahm ES, Sagherian K, Zhu S. Use of Patient Portals in Older Adults: A Comparison of Three Samples. Stud Health Technol Inform. 2016;225:354-8. PMID: 27332221.
[18] Kim EH, Kim,Y. (2010, January). Digital divide: Use of electronic personal health record by different population groups. In 2010 Annual International Conference of the IEEE Engineering in Medicine and Biology (pp. 1759-1762). IEEE.

Challenges of Trustable AI and Added-Value on Health
B. Séroussi et al. (Eds.)

doi:10.3233/SHTI220519

Data Analytics of Electronic Medical Record to Study Racial Diversities in Cardiovascular Diagnosis and Treatment

Maryam PANAHIAZAR[a,b,1], Yorick CHERN[c], Ramon RIOJAS[a], Omar S. LATIF[d], Dexter HADLEY[e] and Ramin E. BEYGUI[a]
[a] *Department of Surgery, Division of Cardiothoracic Surgery, School of Medicine, University of California San Francisco, San Francisco, CA, USA*
[b] *Bakar Computational Health Sciences Institute, School of Medicine, University of California San Francisco, San Francisco, CA, USA*
[c] *Department of Bioengineering, University of California Berkeley, Berkeley, CA, USA*
[d] *Department of Internal Medicine, Mayo Clinic, Rochester, MN, USA*
[e] *Department of Clinical Sciences, College of Medicine, University of Central Florida, Orlando, FL, USA*

Abstract. The study of precision medicine that measures the effects of social, cultural, and environmental influences on health is essential to improve health outcomes. Race is a social concept used historically to divide, track, control populations, and reinforce social hierarchies. Beyond genetics, race is also a surrogate for other socioeconomic factors affecting patient outcomes. Our data analytics study aims to analyze the Electronic Medical Record (EMR) to study patients of different races in diagnosing and treating Coronary Artery Disease (CAD). We found no race discrepancies at the University of California San Francisco Medical Centers. This study opens several new hypotheses for further research in this crucial field.

Keywords. Electronic Medical Record, Racial Study, Cardiovascular Disease, Data Analytics, Precision Medicine

1. Introduction

The risk of developing Coronary Artery Disease (CAD) is markedly different in patients of various races. Precision medicine research promises several approaches to introducing human genomics at the root of racial health disparities. Sometimes, genomic approaches are insufficient to address the various social experiences that often correlate with social behavior. Discrimination in housing and employment, inadequate access to health insurance, and implicit and explicit biases in medical care, for example, substantially impact health outcomes, long-term health, and health disparities. Therefore, precision medicine studies that measure the effects of cultural, social, and environmental influences on health are essential to improve health outcomes. Race is a social concept

[1] Corresponding Author, Maryam PANAHIAZAR; E-mail: Maryam.panahiazar@ucsf.edu

used historically to divide, control populations and reinforce social hierarchies. National Institutes of Health (NIH) promoted the use of racial and ethnic categories created by the US Census Bureau's Office of Management and Budget (OMB) in biomedical research contexts [1,2].They require the new approach toward precision medicine to look at the diversity further than just genomic biomarkers. Racial diversities may help facilitate a deeper understanding of clinical outcomes. This study aims to analyze the Electronic Medical Record (EMR) to find the possible effect of different races on the diagnosis and treatment of CVD.

2. Method

2.1. Study Cohort and Data Selection

The data is from de-identified EMR from 960,129 patients admitted to UCSF during 2011-2018. After authorization to access EMR data for research, the following cohort search criteria were developed for Coronary Artery Disease (CAD), commonly referred to as Ischemic Heart Disease (IHD), based on the ICD10 code (120-125). Patients with missing values specifically for the ICD10 code were excluded. Patients defined as unknown and unspecified were excluded. Patients who met the above criteria led to a cohort size of ~33,000 patients with CAD.

Table 1. It shows the characteristics for a total of patients with CAD diseases (ICD10 code I20-I25).

Characteristics	Percentage	Characteristics	Average
Never Smoked	58.71	Age	69 (y)
Passive Smoke Exposure	0.58	BMI	31.74
Former Smoker	37.8	Blood Pressure (Diastolic)	69.95
Current Every day Smoker	5.25	(Systolic)	130.4
Ethnicity (Hispanic or Latino)	9.42	Cholesterol (Total)	168.65
Ethnicity (Not H or L)	82.71	Cholesterol (LDL)	91.90
Stroke	35.15	Cholesterol (HDL)	52.07
Status (alive)	88.79		
Status (deceased)	10.21		

All data are extracted with MySQL queries from original datasets at UCSF medical center. Characteristics are shown in Table 1. Our data shows five different races in Table 2. The patients whose medical history does not include at least one element from the Current Procedural Terminology (CPT) codes that we defined as various cardiac procedures such as therapeutic and diagnostic catheters, various lab tests, and CABG were eliminated from the initial cohort of patients.

Table 2. It shows the racial distribution for women and men.

Race	Women	Men	Total
Asian	2126	3019	5145
American Indian or Alaska Native	73	110	183
Black or African American	1243	1403	2646
Native Hawaiian or Other Pacific Islander	269	377	646
White or Caucasian	6124	11649	17773

2.2. *Data Analytics Platform and Setup.*

We developed an algorithm that creates a sequence of events and finds the number of days between two events (e.g., the first electrocardiogram (EKG) test and Coronary artery bypass graft surgery (CABG)). Then, we iterated over the entire patient dataset and categorized these numbers of days by each patient's given race. After obtaining the average number of days between different events for each race (e.g., the average number of days between the first EKG and CABG for all races), we performed pairwise 2 sample T-tests corrected with Sidak correction. Lastly, the algorithm was able to take in a threshold value, and this threshold value helped us remove more possible outliers that may skew the results. This threshold represents the maximum number of days allowed in between events. Otherwise, we removed that patient. For example, if a patient has 500 days between their first EKG test and CABG and the threshold is set at 365, this patient is not considered when calculating the average amount of days. The next step was to search for evidence of race-based differences in different race groups. To validate this hypothesis, we determined the first point of an encounter with a physician when a patient was suspected of having CAD. We included the treatment path with suspicion of potential CAD that combined both procedures and medications. Any medication that belongs to the classes of cardiovascular and cardiac drugs, antiplatelets, aspirin, beta-blockers, and statins is included as the starting point for medication. This medication is considered a "first event" for patients suspected of being at risk for CAD. Then, both the men and women datasets are merged on the same paths of the event for each race group defined by experts and ready for hypothesis validation based on p-value < 0.05 and p-value <0.00851 (after Bonferroni Correction) for significant differences. RStudio[2] and Python[3] in the Jupyter[4] notebook have been used for data characteristics and time-series sequence analysis for the path of events and finding possible differences in each race group.

3. Results

Our previous study found significant gender differences in the average waiting days from the first event to diagnostic catheterization [4]. We did not find significant gender-based differences between diagnostic catheterization and therapeutic cardiac catheterization (Percutaneous Coronary Intervention-PCI). We did not find any significant differences between diagnostic catheterization and CABG as well. Since this project requires a pairwise comparison with each race group, we implemented a Bonferroni correction[5] method with the original threshold at $p=0.05$. Therefore, the new threshold set to $p=0.00851$. We added a Bonferroni to reduce the probability of type I errors while we performed multiple sample tests to compare different races. We compared different combinations for more than 70 race group combinations (e.g., American Indian or Alaska Native versus Black or African American). Given this situation, the results are as following. There are no significant differences between different races from first event

[2] https://github.com/rstudio/

[3] https://www.python.org

[4] https://jupyter.org

[5] https://en.wikipedia.org/wiki/Bonferroni_correction

to diagnostic catheterization with both p-value < 0.5 (original threshold) and p-value < 0.00851(after Bonferroni correction). There are no significant differences from the first event to PCI between different races. There are differences between the first event to CABG, as shown in Table 3 between Asian versus American Indian or Alaska Native, American Indian or Alaska Native versus White or Caucasian. From the first event to CABG, Asians waited shorter than other races. White or Caucasians also waited shorter than American Indian or Alaska Natives. There are no significant differences in pairwise comparison for this category after applying correction with the 0.00851 thresholds instead of 0.05 for p-value. According to a study by Garcia et al., the black population in the United States has worse cardiovascular health and higher cardiovascular morbidity and mortality rates than other racial groups [3]. Our study shows that Native Americans or Black have waited longer versus Asian and white to get the invasive procedure from diagnostic catheterization. That could be one of the reasons for worse outcomes in this population. American Indian or Alaska Native also waited longer than any other race groups. Figure 1 shows the average time between each pair of events for different races.

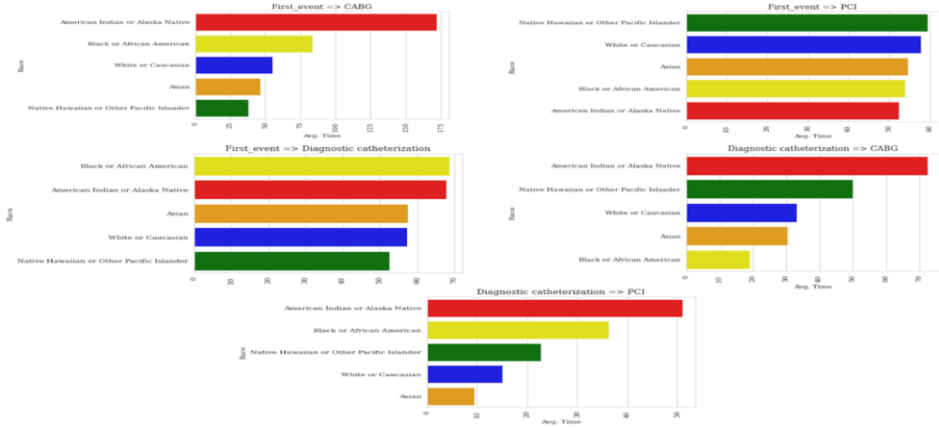


Figure 1. Average waiting days (time) between each event for different races.

Table 3. A comparison between different race groups shows with the p-value <0.05 with 95% confidence interval and p-value < 0.00851 after correction. We show the group of races that are significant with original or corrected threshold. diagnostic catheterization (d-Cath), Versus (v), nonsignificant(nS), Significant (S) Average Waiting Days(AWD), p-value(p)

Pair of event	Race	p, <0.00851, < 0.05	AWD
first-event => CABG	Asian v American Indian or Alaska Native	0.02015, nS, S	46 v 172
	Asian v Black or African America	0.04713, nS, S	46 v 83
	White or Caucasian v American Indian or Alaska Native	0.04044, nS, S	55 v 172
d-Cath=>.PCI	Asian v American Indian or Alaska Native	0.0081, S, S	9 v 51
	Asian v Black or African American	0.0060, S, S	9 v 36
	White or Caucasian v Black or African American	0.0195, nS, S	15 v 36
d-Cath => CABG and first-event =>PCI and first-event =>PCI and first-event => d-Cath	No Significant differences between any races		

4. Discussion

This finding opens several hypotheses, such as considering Black and African Americans sooner to surgery if it is needed or convince them for surgery; look up the social, economic, environmental reason for longer waits. Also finding the comorbidities and other related illnesses for individual races with related social characteristics plus genetics is essential for prevention and a better treatment plan.

5. Limitation

The biggest limitation is considering just EMR for this study. There is an urgent need for considering other resources such as Social Media for racial discrepancies and the possible reasons.

6. Conclusion

Data analytics over UCSF patients who had an invasive procedure makes a few questions, hypotheses, and roadmaps for the future research. For example, what are the possible reasons for the worse outcome in Black or African American population? Maybe staying longer to go for PCI and/or CABG could be one of the reason. Or what is the reason for Asian population for shorter waiting time? This could reflect the effect of culture, environment and social economic in San Francisco BAY Area. More study is needed to confirm this new hypothesis. According to the American college of cardiology, solving racial health disparities is an essential and pressing priority for all in health care. There is an urgent need to address these, both locally and nationally [5]. We plan to extend our data set to UC System data, to study this hypothesis and discover more knowledge for the effect of races in CVD diagnosis and treatment and health outcome.

References

[1] Callier SL. The Use of Racial Categories in Precision Medicine Research. Ethn Dis. 2019 Dec 12;29(Suppl 3):651-658. doi: 10.18865/ed.29.S3.651. PMID: 31889770; PMCID: PMC6919973..

[2] Charkoudian N, Fadel PJ, Robinson AT, Zucker IH. Call for papers on racial differences in cardiovascular and cerebrovascular physiology. Am J Physiol Heart Circ Physiol. 2020 Aug 1;319(2):H249-H250. doi: 10.1152/ajpheart.00524.2020. Epub 2020 Jul 3. PMID: 32618515.

[3] Garcia M, Almuwaqqat Z, Moazzami K, Young A, Lima BB, Sullivan S, Kaseer B, Lewis TT, Hammadah M, Levantsevych O, Elon L, Bremner JD, Raggi P, Shah AJ, Quyyumi AA, Vaccarino V. Racial Disparities in Adverse Cardiovascular Outcomes After a Myocardial Infarction in Young or Middle-Aged Patients. J Am Heart Assoc. 2021 Sep 7;10(17):e020828. doi: 10.1161/JAHA.121.020828. Epub 2021 Aug 25. PMID: 34431313; PMCID: PMC8649258.

[4] Panahiazar M, Bishara AM, Perri J. A Trend of Eight-Years Big Data Analytics of Electronic Medical Records to Review and Study Diagnosis and Treatment of Coronary Artery Disease in Different Genders, *Res. Sq.* (2021) 1–16. doi:https://doi.org/10.21203/rs.3.rs-366930/v1.

[5] Suhendra AD, Asworowati RD, Ismawati T. Racial Differences in Long-Term Cardiovascular Outcomes. *JACCg*. 2020;5:43–54.

Challenges of Trustable AI and Added-Value on Health
B. Séroussi et al. (Eds.)

doi:10.3233/SHTI220520

Local Explanation-Based Method for Healthcare Risk Stratification

Jean-Baptiste EXCOFFIER[a,1], Elodie ESCRIVA[a,b], Julien ALIGON[b] and Matthieu ORTALA[a]
[a] *Kaduceo, Toulouse, France*
[b] *Université de Toulouse-Capitole, IRIT, (CNRS/UMR 5505), Toulouse, France*

Abstract. Decision support tools in healthcare require a strong confidence in the developed Machine Learning (ML) models both in terms of performances and in their ability to provide users a deeper understanding of the underlying situation. This study presents a novel method to construct a risk stratification based on ML and local explanations. An open-source dataset was used to demonstrate the efficiency of this method that well identified the main subgroups of patients. Therefore, this method could help practitioners adjust and build protocols to improve care deliveries that would better reflect patient's risk level and profile.

Keywords. Machine Learning, Explainability, Clustering, Risk Stratification

1. Risk Stratification Method

Clinical decision support tools based on predictive Machine Learning (ML) modeling often provide the prediction alone, raising concerns about their reliability and acceptance. Several methods have thus been perfected to provide local explanations, also called influences, for a individual prediction. Moreover, a recent trend in ML is that predictive tools should not only provide local explanations but also enable the user to contextualize the observation [1]. This need is particularly high in healthcare, where physicians need to link a patient to a more global context so as to deliver the most relevant care. Care management through risk stratification (RS) is thus commonly used in healthcare [2]. It consists in the identification of subgroups, where patients in the same group have similar condition and risk level. Nevertheless, no method has yet been perfected that would take into account both the predictive power of ML models and the identification of such homogeneous subgroups. This study proposes a novel RS procedure based on ML modeling, local explainability methods and clustering that would provide physicians a clearer view of the different profiles of patients, also called subphenotypes or typologies. This method is divided into the three following steps.

Step 1: *Modeling of the underlying dataset using a ML predictive model*, that provides for each patient a risk level, for example a probability of having a disease.

Step 2: *Identification of personal protective and risk factors using a local explainability technique* that quantifies for each patient the influence of each feature.

Step 3: *Identification of subgroups of patients* with respect to their condition and risk level, using a clustering technique on the local explanations computed in the

[1] Corresponding Author, Jean-Baptiste Excoffier: jeanbaptiste.excoffier@kaduceo.com.

previous step. Then a instance selection method can be applied in order to provide clear view of characteristics and specificity of each subgroups.

2. Use-Case

The open-source *Acute Inflamations* dataset was used [3]. It contains information about 120 patients and whether they have an *Acute Inflammation of Urinary Bladder* (AIUB). The techniques used for each step are the following.

Step 1: An ensemble of boosted trees based on the XGBoost implementation was used as the ML predictive risk model.

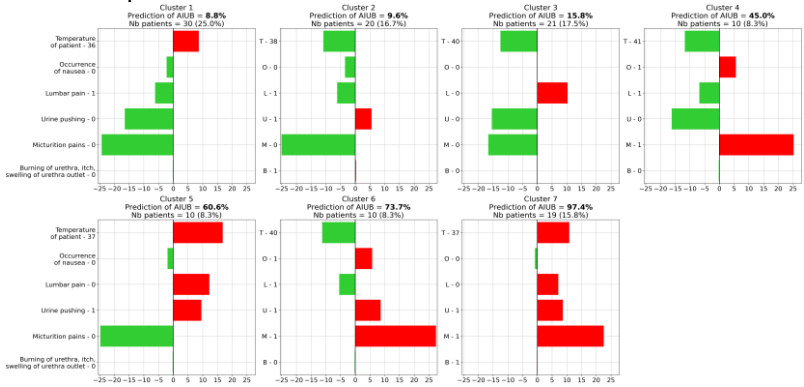


Figure 1. Influences of patients corresponding to the medoid of each identified cluster. Feature names are represented by their initials for subgraph not located on the far-left. Initial values of features are indicated after the hyphen. A positive influence (represented in red) increased the AIUB risk

Step 2: The SHapley Additive exPlanation (SHAP) technique was applied to compute local explanations for each patient [4].

Step 3: The K-Medoid algorithm was used to perform the clustering task on the local explanation values. The optimal number of groups was selected with the Silhouette coefficient. The K-medoid also had the advantage of being both a clustering and instance selection technique, since it can provide for each identified subgroup its most representative patient (*i.e.* medoid), as shown in Figure 1.

Identified subgroups of patients were coherent with medical indications about AIUB given in the original paper [13]. Therefore, it indicates that the novel proposed method is a strong option to consider when the final objective of the data analysis is the construction of a RS, for the purpose of guiding physicians to adjust or create medical protocols that best answer the specific needs of each patient.

References

[1] Lundberg S, Erion G, Chen H, DeGrave A, Prutkin J, Nair B, et al.. From local explanations to global understanding with explainable AI for trees, Nat. Mach. Intell., 2, 56–67; 2020.

[2] Dera JD. Risk stratification: A two-step process for identifying your sickest patients. Family practice management. 2019;26(3):21-6.

[3] Czerniak J, Zarzycki H. Application of rough sets in the presumptive diagnosis of urinary system diseases. In: Artificial intelligence and security in computing systems. Springer; 2003. p. 41-51.

[4] Lundberg S, Lee S. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 2017, vol. 30.

Challenges of Trustable AI and Added-Value on Health
B. Séroussi et al. (Eds.)

doi:10.3233/SHTI220521

A Wide Database for a Multicenter Study on *Pneumocystis jirovecii* Pneumonia in Intensive Care Units

Gabriele DI MECO [a, 1], Sara MORA [a], Daniele Roberto GIACOBBE [b, c, d], Silvia DETTORI [b, c], Ilias KARAISKOS [d, e], Matteo BASSETTI [b, c, d] and Mauro GIACOMINI [a]; on behalf of ESGCIP (Critically Ill Patients Study Group of the European Society of Clinical Microbiology and Infectious Diseases)

[a] *Department of Informatics, Bioengineering, Robotics and System Engineering*
[b] *Department of Health Sciences, University of Genoa, Genoa, Italy*
[c] *San Martino Polyclinic Hospital – IRCCS, Genoa, Italy*
[d] *ESGCIP ((Critically Ill Patients Study Group of the European Society of Clinical Microbiology and Infectious Diseases*
[e] *1st Department of Internal Medicine - Infectious Diseases, Hygeia General Hospital, Athens, Greece*

Abstract. *Pneumocystis jirovecii* pneumonia (PJP) is an opportunistic fungal infection that may affect patients with immunosuppression. In order to improve the diagnosis accuracy for PJP, facilitating the collection of data across Europe to reliably assess the performance of diagnostic tests for PJP is essential to improve the care of critically ill patients developing this severe condition. Such large data can be collected thanks to the contribution of several European hospitals in the compilation of a dedicated electronic Case Report Form (eCRF). The main focus of this work is to create an interface with high ergonomics both in the compilation and in the subsequent validation of the records.

Keywords. Pneumocystis Jirovecii, Pneumonia, eCRF, Relational Database

1. Introduction

Pneumocystis jirovecii pneumonia (PJP) is an opportunistic fungal infection. Classically, PJP has been described in patients with human immunodeficiency virus (HIV) infection, but it may also develop in patients with solid organ transplantation, hematological malignancies and some other pathologies and medical treatments [1-3]. PJP may present as a severe disease with respiratory insufficiency requiring admission to intensive care unit (ICU) and mechanical ventilation. The diagnosis of PJP in ICU (especially in non-HIV patients) is not standardized, and large studies to solidly evaluate the accuracy of diagnostic tools are currently lacking. The aim of this paper is to present a database and a structured system that will be used to collect hundreds of high-quality pseudonymous samples to further improve the early recognition and proper treatment of PJP.

[1] Corresponding Author, Gabriele Di Meco, DIBRIS University of Genoa; E-mail: 4490982@studenti.unige.it.

2. Methods

The study and its electronic Case Report Form (eCRF) was approved by the Liguria ethics committee (305/2021-11538). The protocol requires the collection of features from several fields: Medical History, Hospitalization, Radiology, Microbiology and Laboratory results (specific for PJP or not), Therapy and Outcome. The database structure was developed following the precepts of the Class Diagram in the Unified Modeling Language (UML). More than 20 European centers take part in the data compilation. The structure was designed to support further development, allowing future automatic extraction and storage of congruent data collected by already existing parallel projects like Liguria HIV Network [4], according to the service-oriented approach [5].

3. Results

The designed and developed architecture is composed by a *Blazor* web interface (client), that is able to format and automatically validate the data that will be saved into a *SQL Database* thanks to an *API RESTful service*. The communication and data conversion are managed by *Entity Framework Core*, with a *code first* approach starting from C# Classes. Most fields of the eCRF consist in multiple choice questions, whose values come from a dedicated dataset. The compiler can alternatively indicate a new free-text value, and the revisors will both validate its congruity and evaluate its eventual addition to the dataset.

4. Discussion and Conclusion

The main focus of the project was to create a dedicated eCRF able to guarantee a complete control over the logic relations between its features and their subsequent revision and validation. Providing a similar structure is necessary to facilitate adequate collection and control of data from several centers on a controversial topic which cannot be assessed in single-center or national cohorts: small samples would not provide sufficient power for statistical analyses. The present project may ultimately improve both the prompt recognition and the care of critically ill patients with PJP in European ICUs.

References

[1] Festic E, Gajic O, Limper AH, Aksamit TR. Acute respiratory failure due to pneumocystis pneumonia in patients without human immunodeficiency virus infection: outcome and associated features. Chest 2005; 128:573-579.

[2] Maschmeyer G, Helweg-Larsen J, Pagano L et al. ECIL guidelines for treatment of Pneumocystis jirovecii pneumonia in non-HIV-infected haematology patients. J Antimicrob Chemother 2016; 71:2405-2413.

[3] Giacobbe DR, Cortegiani A, Karaiskos I et al. Performance of Existing Definitions and Tests for the Diagnosis of Invasive Fungal Diseases other than Invasive Candidiasis and Invasive Aspergillosis in Critically Ill, Adult Patients: A Systematic Review with Qualitative Evidence Synthesis. Journal of fungi (Basel, Switzerland) 2021; 7.

[4] Giannini B, et al. From Liguria HIV Web to Liguria Infectious Diseases Network: How a Digital Platform Improved Doctors' Work and Patients' Care. AIDS Research and Human Retroviruses 2018;34: 239-240.

[5] Gazzarata R. et al. A SOA-based platform to support clinical data sharing. Journal of healthcare engineering 2017 (2017). https://doi.org/10.1155/2017/2190679

Challenges of Trustable AI and Added-Value on Health
B. Séroussi et al. (Eds.)

doi:10.3233/SHTI220522

MIMIC-IV as a Clinical Data Schema

Niklas GIESA[a1], Patrick HEEREN[a], Sophie KLOPFENSTEIN[a,b] Anne FLINT[a], Louis AGHA-MIR-SALIM[a], Akira PONCETTE[a], Felix BALZER[a], Sebastian BOIE[a]
[a] *Institute of Medical Informatics, Charité – Universitätsmedizin Berlin*
[b] *Berlin Institute of Health, Charité – Universitätsmedizin Berlin*

Abstract. Routinely collected electronic health records (EHR) in clinical information systems (CIS) are often heterogeneous, have inconsistent data formats and lack of documentation. We use the well-known open-source database schema of MIMIC-IV to address this issue aiming to support collaborative secondary analysis. Over 154 million data records from a German ICU have already been mapped and inserted into the schema successfully. However, discrepancies between the German and US health systems as well as specifics in our clinical source data hinder the direct translation to MIMIC. Evaluating and improving mapping completeness is part of the ongoing research.

Keywords. MIMIC, Electronic Health Records, Critical Care, Clinical Database

1. Introduction

The demand for exchanging standardized Electronic Health Records (EHR) increases with the maturity level of digital innovations in healthcare [1]. To develop and deploy clinical applications, EHR must be stored in a data model which mitigates issues like inconsistencies, format variety or incomplete documentation [1, 2]. The Medical Information Mart for Intensive Care (MIMIC) is a widely used clinical database providing an open-accessible well-documented data schema [3, 4]. Entities for patient characteristics, laboratory values, high-frequency measurements, and billing information are included [4]. Due to its lightweight design, we use MIMIC as a schema to store EHR from ICUs of a German university hospital enabling collaborative secondary analysis. Although previous work used MIMIC for data translation, no known study tried to map data directly from online clinical information systems (CIS) to the schema [5, 6].

2. Methods

The most recently published fourth version of MIMIC was chosen for mapping data fields from clinical MSSQL databases. MariaDB was selected to construct tables by executing the open accessible data definition language (DDL) scripts to create MIMIC-IV [4]. Extract, transform, load (ETL) processes run on Python 3.7.6 and execute SQL statements via the ODBC engine JayDeBe against the source and the target tables. Ethics approval was obtained by an independent ethics committee [7].

[1] Niklas Giesa, Institute of Medical Informatics, Charité – Universitätsmedizin Berlin; Charitéplatz 1 10117 Berlin, E-mail: niklas.giesa@charite.de

3. Results

Up to now, the MIMIC replica includes 37,460 patients with 54,949 ICU stays from March 2019 to October 2021. Eleven tables have been populated with data successfully incorporating more than 154 million EHR. 95% of data fields from the tables *admissions*, *patients* and *transfers* – storing information about the hospital stay – could be mapped to the schema. Mapping completeness ranged from 30 to 70% for *chartevents*, *inputevents*, *procedureevents* and *labevents* holding routinely collected high volatile data. 90% of billing information – stored in *diagnosis_icd* and *procedures_icd* – were transferred to the MIMIC schema; adaptions were necessary because of differences between German and US healthcare systems. 71% of data columns from *d_items* and *d_labitems* could be mapped. In total, 60% (78 out of 129) of data fields have been used for the study.

4. Discussion

Our results illuminate that the openly accessible MIMIC structure can be used to store data from German clinical source systems. However, the transferability of provided columns varies a lot; Paris et. al. state similar issues causing 16% to 80% loss of fields per table [5]. Furthermore, Maier et al. emphasize the lack of mapping concepts for German billing codes, hindering data translation in our study as well [6]. Besides, CIS – designed for daily clinical care – exhibit poor data quality and integrity [1, 2]. In order to provide a curated and interoperable database, further published approaches regarding clinical mapping, data validation and standardization must be evaluated [1, 2, 5, 8].

5. Conclusion

When keeping specifics of source data and health system discrepancies in mind, the schema of MIMIC can be used to efficiently store EHR extracted from German ICUs.

References

[1] Lehne M, Sass J, Essenwanger A, Schepers J, et al. (2019). Why digital medicine depends on interoperability. NPJ digital medicine, 2(1), 1-5.
[2] Nissen F, Quint JK, Morales DR, et al. (2019). How to validate a diagnosis recorded in electronic health records. Breathe, 15(1), 64-68.
[3] Johnson AE, Pollard, TJ, Shen L, et al. (2016). MIMIC-III, a freely accessible critical care database. Scientific data, 3(1), 1-9.
[4] https://mimic.mit.edu/docs/iv/, accessed 15.12.2021 01:09 Berlin, Germany
[5] Paris N, Parrot A (2020). MIMIC in the OMOP Common Data Model. medRxiv, 1-3
[6] Ulrich H, Behrend P, Wiedekopf J, et al. (2021). Hands on the Medical Informatics Initiative Core Data Set—Lessons Learned from Converting the MIMIC-IV. In German Medical Data Sciences 2021, (pp. 119-126). IOS Press. P. 1-5
[7] Ethics vote no. EA1/127/18, chairperson: Prof. R. Morgenstern, ethics committee: Campus Charité - Mitte
[8] Maier C, Lang L, Storf H, et al. (2018). Towards implementation of OMOP in a German university hospital consortium. Applied clinical informatics, 9(01), 054-061

Challenges of Trustable AI and Added-Value on Health
B. Séroussi et al. (Eds.)

doi:10.3233/SHTI220523

Extending the Austrian National EHR System with Patient-Reported Outcome Data

Florian KATSCH[a,d,1], Rada HUSSEIN[a], Stefan SABUTSCH[b], Helene PRENNER[b], Fabian PRASSER[c], Tanja STAMM[d], Georg DUFTSCHMID[d]
[a] *Ludwig Boltzmann Institute for Digital Health and Prevention, Salzburg, Austria*
[b] *ELGA GmbH, Vienna, Austria*
[c] *Berlin Institute of Health at Charité - Universitätsmedizin Berlin, Berlin, Germany*
[d] *Center for Medical Statistics, Informatics and Intelligent Systems, Medical University of Vienna, Vienna, Austria*

Abstract. The Austrian national Electronic Health Record system ELGA is a population-based infrastructure for health data. However, to date, it does not include patient-reported outcomes. In this paper, we describe on-going work on extending ELGA with patient-reported outcome data. This will be done by linking ELGA with the infrastructure of the Health Outcomes Observatory (H2O) initiative. The focus will be on using ELGA's identifier registry for H2O patients and making H2O outcome data accessible in ELGA via an existing ELGA document type for telemonitoring.

Keywords. patient-reported outcomes, H2O, national EHR system, Austria

1. Introduction

The incorporation of patient-reported outcomes (PROs) in routine care is recognized as an important aspect to optimize health systems towards value-based care and to improve patients' care experience and engagement [1]. Europe's current health systems mostly do not promote enough the use of PROs in clinical care [2]. The Health Outcomes Observatory (H2O) initiative [3] will establish an ecosystem to provide PROs for informed health care decision-making to patients, care providers, scientists, and other stakeholders including industry researchers. It will deploy a federated infrastructure for data collection, management, and analysis that will include clinical outcomes as well as PROs, initially in Austria, Germany, the Netherlands and Spain. An upscaling to further European countries is planned later on. In accordance with H2O's goal to offer an integration into national health care ecosystems, the Austrian partners within H2O are exploring how H2O could be linked with Austria's national Electronic Health Record (EHR) system ELGA [4]. First focus points will be the use of (i) the ELGA identifier registry for identification of H2O patients, accessed via Integrating the Healthcare Enterprise (IHE) Patient Demographics Query (PDQ), and (ii) an existing ELGA document type for telemonitoring to make H2O PROs accessible in ELGA, utilizing Health Level Seven (HL7) Clinical Document Architecture (CDA).

[1] Corresponding Author: Florian KATSCH, E-mail: florian.katsch@dhp.lbg.ac.at.

2. Methods

A requirements engineering process will be conducted that accounts for needs of all stakeholder groups. Austria utilizes sector-specific digital personal identifiers (bPK), which prevent cross-sector tracing. The healthcare-specific personal identifier (bPK-GH) is stored in ELGA to support the identification of patients in ELGA-specific tasks. When a new patient is registered in H2O, its identity management component will request the bPK-GH from ELGA via IHE PDQ. ELGA features a telemonitoring document type for storing PROs collected within H2O by means of sensors or questionnaires. We will examine how the H2O dataset can be best fed into the structure of the ELGA document type. This will include a transformation from the HL7 FHIR format used in H2O to the HL7 CDA format used in ELGA. Whereas the FHIR resources and profiles to be used have not yet been defined, the CDA telemonitoring document type has been released by ELGA in 2021. Clinical outcomes and PROs will be stored in a structured and standardized way to enable their effective analysis.

3. Results, Discussion, and Conclusion

The described work is in its early phase. Currently, we assess preferences of different stakeholder groups including organizational and legal requirements for accessing the mentioned identifiers through the H2O infrastructure and storing PROs as telemonitoring documents in ELGA. Another current focus within H2O is the design of the diabetes-specific questionnaires to be used in the first prototype planned for 2022. H2O aims to collect and incorporate PROs along with other health outcomes into healthcare decision making at personal as well as population level and across multiple European health systems. It strives for an integration into existing eHealth infrastructures wherever possible. ELGA provides health data sharing for around 97% of the Austrian population and enables care providers as well as citizens to access cumulative health data. Linking H2O to ELGA will extend the latter's current focus on clinical data with the new dimension of PRO data and at the same time allow H2O to benefit from the well-proven functionalities of an established national EHR system.

References

[1] Basch E. Patient-Reported Outcomes - Harnessing Patients' Voices to Improve Clinical Care. N Engl J Med. 2017 Jan 12;376(2):105–8.
[2] Demiris G, Iribarren SJ, Sward K, Lee S, Yang R. Patient generated health data use in clinical practice: A systematic review. Nurs Outlook. 2019 Apr 26;67(4):311–30.
[3] Stamm T, Bott N, Thwaites R, Mosor E. et al. Building a Value-Based Care Infrastructure in Europe: The Health Outcomes Observatory. NEJM Catalyst Innovations in Care Delivery. 2021;2(3).
[4] Herbek S, Eisl HA, Hurch M, Schator A, Sabutsch St, Rauchegger G, et al. The Electronic Health Record in Austria: a strong network between health care and patients. Eur Surg. 2012 Jun;44(3):155–63.

Challenges of Trustable AI and Added-Value on Health
B. Séroussi et al. (Eds.)

doi:10.3233/SHTI220524

Building a Comprehensive Clinical Data Repository Using FHIR, LOINC and SNOMED

Achim MICHEL-BACKOFEN[a], Thomas PELLIZZARI [c], Jochen ZOHNER[a], Romina BLASINI[b], Kurt MARQUARDT [a]
[a] *University Hospital of Gießen*
[b] *University of Applied Sciences, Giessen*
[c] *AKEDIS GmbH, Innsbruck*

Abstract. In 2018 the University Hospital of Giessen (UHG) moved its hospital information system from an in-house solution to commercial software. The introduction of MEONA and Synedra-AIM allowed for the successful migration of clinical documents. The large pool of structured clinical data has been addressed in a second step and is now consolidated in a HAPI-FHIR server and mapped to LOINC and SNOMED for semantic interoperability in multicenter research projects, especially the German Medical Informatics Initiative (MII) and the Medical Informatics in Research and Care in University Medicine (MIRACUM) consortium.

Keywords. FHIR, LOINC, SNOMED, hospital information system, interoperability

1. Introduction

The University Hospital of Gießen (UHG) operates a HIS on all wards since 1992. In 2018 the in house developed HIS and document archive have been replaced by the MEONA HIS from the German vendor MEONA GmbH and the Synedra-AIM universal archive from Austrian vendor Synedra. To unlock the vast amount of structured clinical data from laboratory, intensive care and anesthesiology for future clinical research, the scope of the Synedra-AIM repository has been widened to include it as a semantic network of FHIR-Resources.

2. Materials and Methods

Structured clinical information within the UHG network resides primarily within 3 main departmental subsystems from laboratory, anesthesiology and intensive care medicine. All systems communicate via HL7 version 2 standard with the Orchestra communication server of UHG. The inherent capability of the FHIR data model to build a semantic network of clinical objects was realized by using HAPI-FHIR server as a clinical data repository (CDR). Together with AKEDIS GmbH we developed an extension of the Synedra-AIM to transform HL7-Version 2 data streams into a HAPI-FHIR CDR. A common structure for business identifiers has been established for the FHIR resources that allows for controlled replication, interoperability and object historization at an

atomic information level. During this transformation, all resources are annotated with LOINC and SNOMED codes for sematic interoperability. For research purposes parts of this CDR are pseudonymized and replicated into project specific HAPI-FHIR servers. Standardized tools transform this data into target environments like i2b2 or csv-files. This allows for fine grained control according to the rules of the european data protection legal framework (Fig. 1).

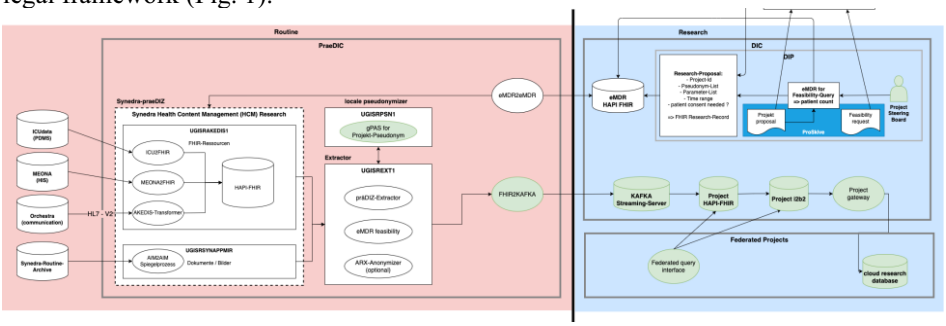


Figure 1: Data flow within the FHIR based research environment at the UHG

3. Results

Starting in 2019 data structures from 4 critical subsystems (HIS, LIS, PDMS and AIMS) have been mapped onto FHIR resources and the clinical data imported into the CDR.

4. Conclusion

Whereas the approach to transform structured clinical data into MII core data set compliant FHIR resources is used within the overall community of the German medical informatics initiative (MII) hospitals, the UHG approach differs in that it builds on an existing HL7-V2 infrastructure without changes to existing subsystems and utilizes the FHIR CDR for both the production and research environment. The benefits have been realized within multicenter research frameworks like the MII and MIRACUM [1].

References

[1] Kapsner LA, Kampf MO, Seuchter SA, Gruendner J, Gulden C, Mate S, et al. Reduced Rate of Inpatient Hospital Admissions in 18 German University Hospitals During the COVID-19 Lockdown. Front Public Health. 2021 Jan 13;8:594117. doi: 10.3389/fpubh.2020.594117. PMID: 33520914; PMCID: PMC7838458.

Challenges of Trustable AI and Added-Value on Health
B. Séroussi et al. (Eds.)

doi:10.3233/SHTI220525

A Deep Learning Method for Automatic Identification of Drusen and Macular Hole from Optical Coherence Tomography

Teresa PACE[a,1], Nicolò DEGAN[a], Rosa GIGLIO MD[b], Daniele TOGNETTO MD[b] and Agostino ACCARDO[a]
[a] *Department of Engineering and Architecture, University of Trieste, Trieste, Italy*
[b] *Eye Clinic, University of Trieste, Trieste, Italy*

Abstract. Deep Learning methods have become dominant in various fields of medical imaging, including ophthalmology. In this preliminary study, we investigated a method based on Convolutional Neural Network for the identification of drusen and macular hole from Optical Coherence Tomography scans with the aim to assist ophthalmologists in diagnosing and assessing retinal diseases.

Keywords. Deep Learning, Convolutional Neural Networks, Optical Coherence Tomography, Drusen, Macular Holes, Retinal diseases

1. Introduction

Deep Learning (DL) systems have been demonstrated to have an important role in image-centric specialties, such as radiology, dermatology, pathology, and ophthalmology [1]. In ophthalmology, the adoption of DL methods to Optical Coherence Tomography (OCT) has shown promising results for the identification of retinal diseases [2-3]. In this work, therefore, we used a simple and efficient Convolutional Neural Network (CNN) to identify automatically drusen, which characterizes age-related macular degeneration, and macular hole (MH) from OCT images. The method forms the basis for the development of a diagnostic support system that can assist ophthalmologists during the diagnosis and assessment of retinal diseases.

2. Methods

OCT scans were extracted from the public dataset "Retinal OCT - C8" [4] and cleaned appropriately. The final dataset consisted of 9000 images randomly divided into training (70%), validation (20%) and test (10%) sets keeping class balance.

The classification of OCT images as normal, drusen or MH was performed using a VGG16 CNN network. The network was pretrained on the ImageNet dataset and fine-tuned. Data pre-preprocessing included normalization and data augmentation (horizontal flip, random brightness, and random rotation). The model was trained for 38 epochs

[1] Corresponding Author, Teresa Pace, Dipartimento di Ingegneria e Architettura, Università degli Studi di Trieste, Via Alfonso Valerio 6/1, 34127 Trieste (TS), Italy; E-mail: teresa.pace@phd.units.it.

(early stopping) in batches of 64 samples using Adam optimizer with a learning rate of 10^{-4} and binary cross entropy loss.

3. Results

The performances of the model were evaluated on the test set by computing confusion matrix (figure 1) and receiver operating characteristic (ROC) curves (figure 2). The system achieved kappa coefficient $\kappa = 0.92$ and macro-average AUC = 0.96. Gradient-weighted Class Activation Mapping (Grad-CAM) was used to provide visual explanation heatmaps for results interpretation (figure 3).

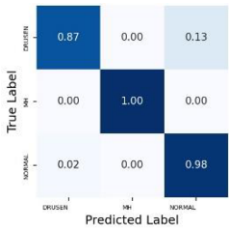


Figure 1. Confusion matrix computed on the test set.

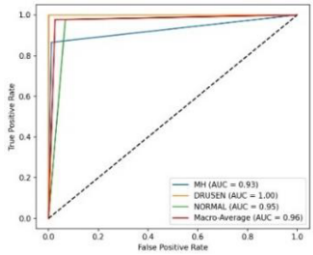


Figure 2. ROC curves for normal, drusen and MH classification.

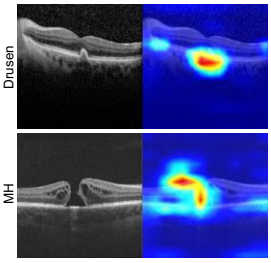


Figure 3. Visual explanation generated by Grad-CAM.

4. Discussion

The proposed method successfully identifies drusen and MH from OCT scans, reaching high classification accuracy. Kappa coefficient and macro-average AUC demonstrate that there is a high agreement between classification and ground truth. Grad-CAM images confirm that the DL model identifies characteristic features of drusen and MH.

5. Conclusions

Our preliminary study shows that a method based on DL was effective at distinguishing normal OCT from OCT presenting diseases features, identifying drusen and MH with high accuracy. This can lead to the development of a more sophisticated diagnostic support system for the diagnosis and assessment of retinal diseases.

6. References

[1] Ting DSW, Peng L, Varadarajan AV, Keane PA, Burlina PM, Chiang MF, Schmetterer L. Deep learning in ophthalmology: The technical and clinical considerations. Prog Retin Eye Res. 2019 Sep;72:100759.

[2] Choi KJ, Choi JE, Roh HC, Eun JS, Kim JM, Shin YK, Kang MC, Chung JK. Deep learning models for screening of high myopia using optical coherence tomography. Sci Rep. 2021 Nov;1:21663.

[3] Ye X, Wang J, Chen Y, Lv Z, He S. Automatic screening and identifying myopic maculopathy on optical coherence tomography images using deep learning. Transl Vis Sci Technol. 2021 Nov;10(13):10.

[4] Naren, OS. Retinal OCT-C8 [internet]. Kaggle; 2021 Oct 5 [updated 2021 Oct 24; cited 2022 Jan 21]. Available from: https://www.kaggle.com/obulisainaren/retinal-oct-c8.

Challenges of Trustable AI and Added-Value on Health
B. Séroussi et al. (Eds.)

doi:10.3233/SHTI220526

The Use of Machine Learning Techniques to Predict Diabetes in Patients with Cystic Fibrosis

Tajda BOGOVIČ[a], Peter KOKOL[a], Jernej ZAVRŠNIK[b] and Helena BLAŽUN VOŠNER[b]
[a] *Faculty of Electrical Engineering and Computer Science, University of Maribor, Slovenia*
[b] *Community Healthcare Center dr. Adolf Drolc, 2000 Maribor, Slovenia*

Abstract. The accuracy of the prognosis of diabetes in patients with cystic fibrosis is crucial, as it highly connected with mortality and other complications. The prognosis of diabetes is a time-consuming process. Usually, it is performed by medical staff and can often lead to misdiagnosis. The aim of the study was to analyze and evaluate risk factors of developing diabetes in patients diagnosed with Cystic Fibrosis by using classification machine learning techniques. The ECFS data register was used to train and test the models. Visualization of our results using SHAP values highlights that most important features are age, antibiotic treatment, FEV1 value and lung transplant as risk predictors for diabetes.

Keywords. Machine learning, SHAP Values, Diabetes, Cystic fibrosis

1. Introduction

Cystic fibrosis (CF) is the most common recessive inherited disease among Caucasians. More than 70.000 people worldwide are diagnosed with CF. CF can cause various health problem, such as liver disease, diabetes, inflammation of the pancreas, and kidney stones [1]. Cystic fibrosis related diabetes (CFRD) is the most common complication of CF, affecting at least half of the adult population [2]. It has an impact on decreasing lung function and increasing the rate of morbidity and mortality. Early identification and treatment is crucial, as the symptoms are usually not immediately recognized. Due the lack of studies focusing on predicting risk factors for diabetes in patients with cystic fibrosis, we used different classification models for the ECFS (European Cystic Fibrosis Society) dataset to predict risk factors for this specific population.

2. Methods

ECFS dataset includes data from more than 49,000 people with CF, from 38 participating countries, and longitudinal data from 2008 to 2018. The database contains data on 389 555 entries of 55547 unique patients. The registry comprises a list of annual follow-up variables for individual CF patients that includes demographics, genetic mutations, airway colonization and microbiological infections, comorbidities and complications, transplantation, hospitalization, spirometry and therapeutic management. While machine

learning has been already used in CF [3] we developed our own optimized approach that was used on the whole ECFS dataset using the steps bellow:
Step 1: The process started from the data collection which was preprocessed.
Step 2: The k-NN approach was applied for data imputation only on features with an acceptable percentage of missing values (≤ 30%) to increase the completeness of the data.
Step 3: Synthetic Minority Oversampling Technique was used to balance the data.
Step 4: A total of 43 attributes were divided into training and test data set in the ration 70:30 and importance of features were explained using SHAP (Shapely Additive Explanations) values to make results easier to understand for medical professionals [4].

3. Results and Discussion

Table 1, presents the Accuracy and Area under ROC curve (AUC) (for three popular machine learning models and Figure 1 the SHAP values of the best model (Cat Boost).

Table 1. Models' accuracy and AUC

Model	Accuracy [%]	AUC [%]
Catboost Model	91±0.06	92±0.01
Decision Tree Classifier	81±0.01	81±0.02
Random Forest Classifier	84±0.02	83±0.03

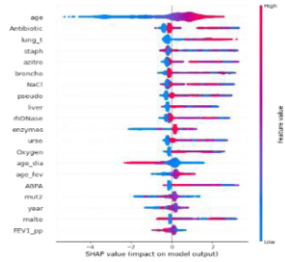


Figure 1. Cat Boost Model - SHAP values

Cat Boost classification indicates that most important features in patients with both diabetes and CF are age, antibiotic treatment, and lung transplant, as well as FEV1 (Volume of air blown). Same attributes were also most prolific in other models. Our study also highlights the advantage of presenting the results of machine learning with SHAP values which can be also easily understood by non-machine learning experts in our case physicians and other health professionals.
Funding: The research was partially funded by EC H2020 Project STAMINA- GA 883441

References

[1] De Boeck K. Cystic fibrosis in the year 2020: A disease with a new face. Acta paediatrica. 2020 May;109(5):893-9. doi: 10.1111/apa.15155.
[2] Iafusco F, Maione G, Rosanio FM, Mozzillo E, Franzese A, Tinto N. Cystic Fibrosis-Related Diabetes (CFRD): Overview of Associated Genetic Factors. Diagnostics. 2021 Mar;11(3):572.
[3] Alaa AM, van der Schaar M. Prognostication and Risk Factors for Cystic Fibrosis via Automated Machine Learning. Sci Rep. 2018 Jul 26;8(1):11242.
[4] Lu S, et al. Understanding Heart Failure Patients EHR Clinical Features via SHAP Interpretation of Tree-Based Machine Learning Model Predictions. In AMIA An Symp Proceedings 2021 (Vol. 2021, p. 813).

Challenges of Trustable AI and Added-Value on Health
B. Séroussi et al. (Eds.)

doi:10.3233/SHTI220527

Wake-up Stroke Outcome Prediction by Interpretable Decision Tree Model

Miloš AJČEVIĆ[a,1], Aleksandar MILADINOVIĆ[a], Giovanni FURLANIS[b], Alex BUOITE STELLA[b], Marcello NACCARATO[b], Paola CARUSO[b], Paolo MANGANOTTI[b] and Agostino ACCARDO[a]
[a] *Department of Engineering and Architecture, University of Trieste, Trieste, Italy*
[b] *Department of Medicine, Surgery and Health Science, University of Trieste, Trieste, Italy*

Abstract. Outcome prediction in wake-up ischemic stroke (WUS) is important for guiding treatment strategies, in order to improve recovery and minimize disability. We aimed at producing an interpretable model to predict a good outcome ($NIHSS_{7\text{-}day}$<5) in thrombolysis treated WUS patients by using Classification and Regression Tree (CART) method. The study encompassed 104 WUS patients and we used a dataset consisting of demographic, clinical and neuroimaging features. The model was produced by CART with Gini split criterion and evaluated by using 5-fold cross-validation. The produced decision tree model was based on NIHSS at admission, ischemic core volume and age features. The predictive accuracy of model was 86.5% and the AUC-ROC was 0.88. In conclusion, in this preliminary study we identified interpretable model based on clinical and neuroimaging features to predict clinical outcome in thrombolysis treated wake-up stroke patients.

Keywords. Wake-up stroke, Predictive modeling, Clinical outcome, Classification and Regression Tree

1. Introduction

Ischemic stroke is nowadays highly treatable with thrombectomy and intravenous thrombolysis reperfusion treatments, also in case of wake-up stroke (WUS) [1]. Outcome prediction for acute ischemic stroke treatment is still challenging and only a few studies focused on WUS. National Institutes of Health Stroke Scale (NIHSS) measured on the 7th day after the ischemic event ($NIHSS_{7\text{-}day}$) can be used as an outcome measure of acute ischemic stroke treatment [2]. The Classification and Regression Tree (CART) models provide better interpretability and practical usability important especially in emergency setting such as ischemic stroke. This preliminary study aimed at producing an interpretable model to predict a good outcome ($NIHSS_{7\text{-}day}$<5) in thrombolysis treated WUS patients by CART.

[1] Corresponding Author, Miloš Ajčević, Department of Engineering and Architecture, University of Trieste, Via A. Valerio 10, 34127 Trieste, Italy; E-mail: majcevic@units.it.

2. Methods

The study was conducted on 104 WUS patients admitted to the Stroke Unit of the University Medical Hospital of Trieste, Italy, who underwent thrombolysis treatment. In this study, we used a dataset consisting of 27 demographic, clinical and neuroimaging features. Good outcome class was defined with $NIHSS_{7\text{-day}}<5$, while bad outcome class with $NIHSS_{7\text{-day}} \geq 5$. The model was produced by CART with Gini split criterion. The model was evaluated by using 5-fold cross-validation.

3. Results

The produced decision tree model is reported in Figure 1a. The model is based on NIHSS at admission, ischemic core volume and age features. The evaluated model accuracy was 86.5%, the confusion matrix is reported in Figure1b. The area under the ROC curve (AUC) was 0.88.

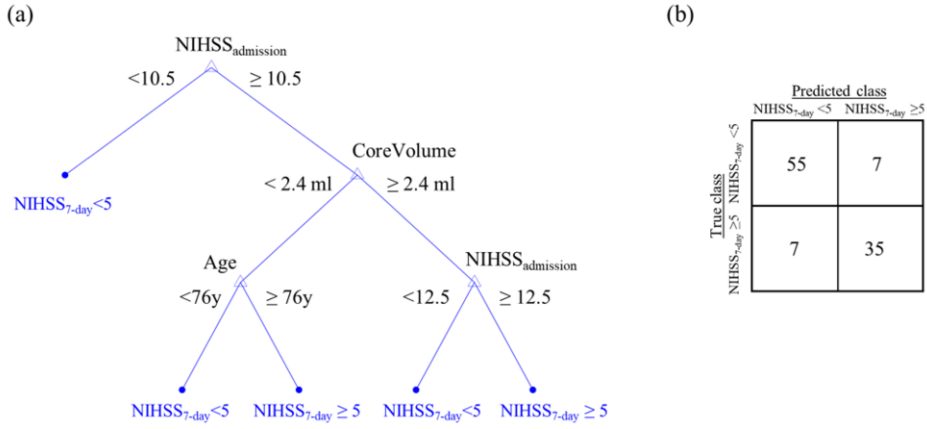


Figure 1. (a) The produced decision tree model (b) Confusion matrix.

4. Conclusions

In conclusion, in this preliminary study we identified an interpretable model to predict clinical outcome in thrombolysis treated wake-up stroke patients. If confirmed in a larger sample size, these findings provide an interpretable tool for early post-stroke prognosis, which is essential to guide therapeutic and rehabilitation strategies.

References

[1] Furlanis G, M. Ajčević M, Stella AB, et al. Wake-up stroke: thrombolysis reduces is-chemic lesion volume and neurological deficit. Journal of neurology. 2020; 267(3):666-673.

[2] Chalos V, van der Ende NAM, Lingsma HF, et al. National Institutes of Health Stroke Scale: An Alternative Primary Outcome Measure for Trials of Acute Treatment for Ischemic Stroke. Stroke. 2020; 51(1):282-290.

Challenges of Trustable AI and Added-Value on Health
B. Séroussi et al. (Eds.)

doi:10.3233/SHTI220528

How Does Triage by an Electronic Symptom Checker Match with Triage by a Nurse?

Tuomas KOSKELA[a], Ville LIU[b], Minna KAILA[c]
[a] *Tampere University, Finland*
[b] *University of Helsinki, Finland*
[c] *The Finnish Medical Society Duodecim, Finland*

Abstract. Omaolo© electronic symptom checkers (ESCs) have been developed to make triage for primary health care patients in Finland. Based on the analysis of the patient's responses to a set of questions, the ESC classifies him/her as emergent, urgent, not urgent, or advices on self-care. In this study the user answered the questions posed by the electronic symptom checker, after which a nurse assessed the urgency of the same user's symptom. The triage nurse was not allowed to know the result of the electronic symptom assessment until he or she had assessed the patient's condition. The level of triage was compared between ESC and nurse in each individual case. Findings from 825 individual cases were analyzed. The mean "exactly matched" for all symptom estimates was 52.6%. The mean "exactly matched" or "overconservative but suitable" for all symptom assessments was 66.6%. Safe assessments of electronic symptom checkers accounted for 98.6% of all assessments. A case was defined as "safe" if the recommendation for action given by the symptom assessment was at most one level less urgent than the nurse's triage assessment of the same case. The findings show that electronic symptom assessments are safe compared to the assessment of an experienced nurse

Keywords. triage, symptom checker, eHealth, primary health care

1. Introduction

In a systematic review no electronic symptom checker outperformed general practitioner in diagnostic accuracy or in safety of urgency advice [1]. The performance on appropriate triage varies notably between apps [1,2]. In addition, electronic symptom checkers, on average, make triage sensitively to the need for more urgent care than the user would actually need [2,3,4]. Omaolo© electronic symptom checkers (ESCs) have been developed to make triage for primary care patients in Finland. Traditionally triage has been made by a nurse in Finland.

Based on the analysis of the patient's responses to a set of questions, the ESC classifies him/her as emergent, urgent, not urgent, or advices on self-care. The user answers questions about their symptoms on the Omaolo© website and gets triage guidance. The idea is to help the users to more adequately assess their condition, and to ease the professionals' triage workload.

In previous studies clinical validation of electronic symptom checkers has been made by clinical case vignettes [1,2]. In this study it will be used the real life users in a real life setting.

2. Methods

The study was a multicenter study in 14 primary care organizations. The user answered the questions posed by the electronic symptom checker, after which a nurse assessed the urgency of the same user's symptom. The triage nurse was not allowed to know the result of the electronic symptom assessment until he or she had assessed the patient's condition. Data were collected in 2019-2020. The level of triage was compared between ESC and nurse in each individual case.

3. Results

Findings from 825 individual cases were analyzed. The mean “exactly matched” for all symptom estimates was 52.6%. The mean "exactly matched" or "overconservative but suitable” for all symptom assessments was 66.6%. Safe assessments of electronic symptom checkers accounted for 98.6% of all assessments. A case was defined as "safe" if the recommendation for action given by the symptom assessment was at most one level less urgent than the nurse's triage assessment of the same case.

4. Conclusions

The findings show that Omaolo electronic symptom assessments are safe compared to the assessment of an experienced nurse.

References

[1] G Gilbert, Stephen et al. How accurate are digital symptom assessment apps for suggesting conditions and urgency advice? A clinical vignettes comparison to GPs. BMJ Open 2020;10:e040269
[2] Semigran HL, Linder JA, Gidengil C, Mehrotra A. Evaluation of symptom checkers for self diagnosis and triage: audit study BMJ 2015;351:h3480
[3] Powley, L., McIlroy, G., Simons, G., & Raza, K. Are online symptoms checkers useful for patients with inflammatory arthritis?. BMC musculoskeletal disorders 2016;17:362.
[4] Poote AE, French DP, Dale J, Powell J. A study of automated self-assessment in a primary care student health centre setting. J Telemed Telecare 2014;20:123-7.

Challenges of Trustable AI and Added-Value on Health
B. Séroussi et al. (Eds.)

doi:10.3233/SHTI220529

Using Explainable Artificial Intelligence Models (ML) to Predict Suspected Diagnoses as Clinical Decision Support

Zully RITTER[a,1], Stefan VOGEL[a], Frank SCHULTZE[b], Kerstin PISCHEK-KOCH[a], Wiebke SCHIRRMEISTER[c], Felix WALCHER[c,e], Rainer RÖHRIG[d,e], Tibor KESZTYÜS[a], Dagmar KREFTING[a] and Sabine BLASCHKE[b,e]

[a] *Institute of Medical Informatics, University Medicine Göttingen, Georg-August University, Göttingen, Germany*

[b] *Central Emergency Department, University Medicine Göttingen, Georg-August University, Göttingen, Germany*

[c] *Department of Trauma Surgery, Otto-von-Guericke University Magdeburg, Magdeburg, Germany*

[d] *AKTIN-Research Group, Germany*

[e] *Institute of Medical Informatics, Medical Faculty of RWTH Aachen University, Aachen, Germany*

Abstract. The complexity of emergency cases and the number of emergency patients have increased dramatically. Due to a reduced or even missing specialist medical staff in the emergency departments (EDs), medical knowledge is often used without professional supervision for the diagnosis. The result is a failure in diagnosis and treatment, even death in the worst case. Secondary: high expenditure of time and high costs. Using accurate patient data from the German national registry of the medical emergency departments (AKTIN-registry, Home - Notaufnahmeregister (aktin.org)), the most 20 frequent diagnoses were selected for creating explainable artificial intelligence (XAI) models as part of the ENSURE project (ENSURE (umg.eu)). 137.152 samples and 51 features (vital signs and symptoms) were analyzed. The XAI models achieved a mean area under the curve (AUC) one-vs-rest of 0.98 for logistic regression (LR) and 0.99 for the random forest (RF), and predictive accuracies of 0.927 (LR) and 0.99 (RF). Based on its grade of explainability and performance, the best model will be incorporated into a portable CDSS to improve diagnoses and outcomes of ED treatment and reduce cost. The CDSS will be tested in a clinical pilot study at EDs of selected hospitals in Germany.

Keywords. Clinical Decision Support, Explainable Artificial Intelligence, Machine Learning, Diagnoses Prediction, Emergency Department

1. Introduction

The number and complexity of emergency cases in emergency departments (EDs) have increased substantially in Germany. High pressure for correct diagnoses in limited time and the reduced number of specialized medical staff in EDs are catalysts leading to

[1] Corresponding Author, Zully Ritter, University Medicine Göttingen, Georg-August University, Institute of Medical Informatics, D-37075, Germany; E-mail: zully.ritter@med.uni-goettingen.de.

erroneous diagnoses. Frequently unproven medical knowledge used for diagnosis identification leads to diagnosis failure or even death in the worst scenery. The German interdisciplinary research project ENSURE (grant ZMVI 2520DAT803) addressed this problem (ENSURE (umg.eu)). The aim is to develop an explainable artificial intelligence machine learning model, to predict suspected diagnoses for patients in the ED. The nationwide collected ED-patient data from the AKTIN registry has been used. The diagnoses suggested by the XAI are to be used as support decisions. Thus, the patient`s diagnosis and treatment are to be performed only by the medical doctor.

2. Materials and Methods

On the Data: The AKTIN-Data registry from ED patients (2017-2020) was used to create explainable ML models (eXML) [1]. The most 20 frequent diagnoses were used to train (80%) and test the eXML-models. 137.152 samples matching the target diagnoses and 51 features (vital signs and symptoms), and 1 column with the associated diagnosis (ICD-10 codes) were analyzed. *On the eXML-Models*: Different eXML models were created and compared. Performance was evaluated using a multiclass confusion matrix, ROC curves, and predictive accuracy. Shap was used for explainability. Sensitivity was preferred over specificity. These processes were already tested using open data [2].

3. Results

Following ICD-10 codes were the most frequent diagnoses in Germany: (I21, I26, J44, I10, I20, E87, I63, S06, K56, J18, F10, I48, F10, I48, E11, G45, T78, R55, A41, K80, I50, J16, M54, J15, S72, A40, J17). These diagnoses were confirmed by ED experts matching with those selected in ENSURE. The best values from the eXML logistic regression (LR) and random forest model (RFM) were: predictive accuracy of 0.927 (LR) and 0.99 (RFM). The area under the curve (AUC, “one vs. rest (ovr)” mean) was: 0.98 (LR) and 0.99 (RFM). Feature reduction, grid search, and data balance using data sets of the E.Care ED within the clinical study centers will be made in future analysis.

4. Conclusions

1) Secondary ED data from the German AKTIN-registry database was proven usable to develop explainable machine learning models. 2) Physiological filters for vital signs and symptoms variables were a key factor for achieving high accuracy and grade of explainability. 3) Random Forest, due to the higher accuracy, AUC, and explainability, is recommendable to be integrated into a CDSS for diagnosis prediction in ED.

References

[1] Gilpin LH, Bau D, Yuan BZ, Bajwa A, Specter M, Kagal L. Explaining Explanations: An Overview of Interpretability of Machine Learning. In: 2018 IEEE 5th International Conference on data science and advanced analytics (DSAA). Turin, Italy; 2018. p. 80–9.
[2] https://share.streamlit.io/zritter2050/cormeum/main/hrfApp.py

Challenges of Trustable AI and Added-Value on Health
B. Séroussi et al. (Eds.)

doi:10.3233/SHTI220530

The Prediction of Fall Circumstances Among Patients in Clinical Care – A Retrospective Observational Study

Sven REHFELD[a], Matthias SCHULTE-ALTHOFF[b], Fabian SCHREIBER[b], Daniel FÜRSTENAU[a,b,c], Anatol-Fiete NÄHER[b,d,1], Armin HAUSS[b], Charlotte KÖHLER[a] and Felix BALZER[b]
[a] *Department of Information Systems, Freie Universität Berlin, Germany*
[b] *Institute of Medical Informatics, Charité – Universitätsmedizin, Germany*
[c] *Department of Digitalization, Copenhagen Business School, Denmark*
[d] *Data Management Unit, Robert Koch Institute, Germany*

Abstract. Standardized fall risk scores have not proven to reliably predict falls in clinical settings. Machine Learning offers the potential to increase the accuracy of such predictions, possibly vastly improving care for patients at high fall risks. We developed a boosting algorithm to predict both recurrent falls and the severity of fall injuries. The model was trained on a dataset including extensive information on fall events of patients who had been admitted to Charité – Universitätsmedizin Berlin between August 2016 and July 2020. The data were recorded according to the German expert standard for fall documentation. Predictive power scores were calculated to define optimal feature sets. With an accuracy of 74% for recurrent falls and 86% for injury severity, boosting demonstrated the best overall predictive performance of all models assessed. Given that our data contain initially rated risk scores, our results demonstrate that well trained ML algorithms possibly provide tools to substantially reduce fall risks in clinical care settings.

Keywords. fall prediction, machine learning, clinical care, retrospective study

1. Introduction

Falls and their physical and psychological consequences represent an omnipresent danger for elderly people. Fall risk assessment tools, such as the highly used 'Morse Falls Scale' assessment tool have limited predictive power for fall risk [1]. Consequently, numerous approaches to deploy data-driven methods have been deployed within the last years. Yet, most papers focus exclusively on fall risk prediction, considering categories, such as demographic data [2,3], medication [2,4], pre-existing conditions [2,5], mobility [6] and leave fall circumstances unaddressed. Fall circumstances refer to context information available in relation to a fall, such as the location, cause, or injury.

[1] Corresponding Author, Anatol-Fiete Näher; E-mail: anatol-fiete.naeher@charite.de.

2. Methods

We deployed and evaluated a boosting algorithm to predict the probability of repeated falls and the severity of fall injuries. The present dataset from the university hospital Charité Berlin comprises a total of 8,874 reported fall accidents, each with 121 different reporting features, covering a total of 6,424 patients. We accounted for unequal distribution of the dependent variable use Synthetic Minority Oversampling Technique (SMOTE) for synthetic data generation. A range of features for fall events and medication were constructed using natural language processing. Two different prediction models for different fall circumstances are developed throughout this paper: prediction of multiple falls and the severity of fall injuries. By specifying a logistic regression, a random forest, a neural network and a support vector machine as alternate classifiers, we evaluated the predictive performance of the boosting model.

3. Results, Discussion and Conclusion

The results of our study show that the severity of injuries and multiple falls can be predicted with an accuracy of up to 86% and 74%, respectively, using a XGBoost classifier. Furthermore, we are able to show the most common fall circumstances for multiple falls, which include gait disorders or gait aids. Furthermore, the number of inpatient days and whether the patient was visiting the toilet have a significant influence. For the second model, it was shown that the severity of falls is significantly associated to hospital days, gender, but also to time of day and season. In addition, neurosurgical patients in particular appear to be at risk of more severe fall injuries. We extracted 104 features from the ambulant care dataset but only a small fraction of features have proven to usefully predict fall severity, while most features only contribute little or nothing. Relevant factors include demographic (Lee et al., 2020), inpatient (Hsu et al., 2020), or medication data (Beauchet et al., 2020) as well as assessment tools such as the GCS score or the NANDA classification for patient mobility [7]. We conclude that the objective of this work was to develop different models for predicting fall circumstances. The findings of this work can serve as a basis for further research in the area of fall circumstance prediction, offering improvements in preventive care interventions.

References

[1] Kim EA, Mordiffi SZ, Bee WH, Devi K, Evans D. Evaluation of three fall-risk assessment tools in an acute care setting. J Adv Nurs. 2007 Nov;60(4):427-35.
[2] Ye C, et al. Identification of elders at higher risk for fall with statewide electronic health records and a machine learning algorithm. Int J Med Inform. 2020 May;137:104105.
[3] Wang L, et al. Preventing inpatient falls with injuries using integrative machine learning prediction: a cohort study. NPJ Digit Med. 2019 Dec 12;2:127.
[4] Hsu YC, Weng HH, Kuo CY, Chu TP, Tsai YH. Prediction of fall events during admission using eXtreme gradient boosting: a comparative validation study. Sci Rep. 2020 Oct 8;10(1):16777.
[5] Kuspinar A, Hirdes JP, Berg K, McArthur C, Morris JN. Development and validation of an algorithm to assess risk of first-time falling among home care clients. BMC Geriatr. 2019 Oct 14;19(1):264.
[6] Beauchet O, et al. Falls Risk Prediction for Older Inpatients in Acute Care Medical Wards. J Nutr Health Aging. 2018;22(1):131-137.
[7] Young AJ, et al. Using Machine Learning to Make Predictions in Patients Who Fall. J Surg Res. 2021 Jan;257:118-127.

Challenges of Trustable AI and Added-Value on Health
B. Séroussi et al. (Eds.)

doi:10.3233/SHTI220531

A Log Analysis Exploring the Predictors of Electronic Health Record Access by Clinicians for Consumers Aged ≥65 Who Present to the Emergency Department

Alexandra MULLINS[a,b,1], Helen SKOUTERIS[c,d], Heather MORRIS[c] and Joanne ENTICOTT[a]
[a] *Monash Centre for Health Research and Implementation, School of Public Health, Monash University, Victoria, Australia*
[b] *Cabrini Health, 181-183 Wattletree Rd, Malvern, VIC, Australia*
[c] *Health and Social Care Unit, School of Public Health, Monash University, Victoria, Australia*
[d] *Warwick Business School, University of Warwick, Coventry, UK*

Abstract. Electronic health records are widely implemented, yet little is understood around adoption and use in the ED setting. Older patients (≥65) are a cohort likely to benefit from use. The primary outcome (MHR access) was explored using logistic regression of 9 independent variables. 28.33% of patients had their MHR accessed within 3 days of presenting. Access is more likely when patients arrive via urgent ambulance and/or are triaged as critical.

Keywords. electronic health record, My Health Record, emergency department, medical system, hospital

1. Introduction

Older patients frequently present to the emergency department (ED), often presenting with many comorbidities that can complicate the delivery of care [1]. Solutions to support improved quality of care and patient outcomes in the ED are highly sought after, and present an opportunity to deliver great benefit to older patients. Electronic health records (EHRs), digital shared versions of a patient health record, are one of the proposed solutions [2]. In Australia, the national personally controlled EHR (known as My Health Record, MHR) was implemented as an opt-out system in 2019. Despite the potential, the impact, use and predictors of MHR use in the ED has not yet been explored [3]. Based on previous research [4], we hypothesize: increasingly age, greater triage urgency and arrival via ambulance would be associated with MHR access by ED clinicians for patients who present to the ED aged over 65.

[1] Corresponding author, Alexandra Mullins, School of Public Health and Preventative Medicine, Monash UniversityAustralia; E-mail: Alexandra.mullins@monash.edu.

2. Methods

Secondary routinely-collected log data was explored retrospectively, linked to attendance data (of patients 65 years and older) and human resources data. The log encompassed all patients who presented to the ED between August 2019-2021 at Cabrini Health (Melbourne, Australia). The primary outcome was MHR access (or not) by an ED nurse, doctor or pharmacist within 3 days of ED presentation. Stata V6 facilitated logistic regression of 9 independent variables (age, gender, triage category, presentation time, arrival method, gender, referral type, length of stay (LOS) and admission.

3. Results

A total of 22,510 patients 65 years of age, or older, presented to the ED at the study site during the study period. Of these presentations, a total of 28.33% patients (n=6,377) had their MHR accessed within 3 days of presenting. Pharmacists, doctors, and nurses accessed 25.77%, 4.07% and 0.64% of patient MHRs (respectively).

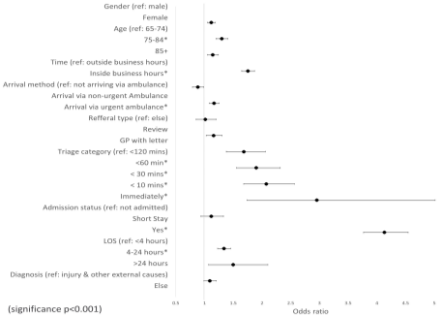


Figure 1. Forest plot, predictors of MHR access within 3 days of presentation to the ED.

4. Discussion

There is low prevalence of MHR access (for patients aged over 65) in the ED, particularly by doctors and nurses. Albeit, pharmacists appear to see benefit in accessing the MHR system – such as when patients arrive via urgent ambulance and/or are triaged as critical. The efficiency gain in this situation should be explored, and leveraged to promote use of the system if and where it makes sense. This research emphasizes the need for policy targeting MHR use by clinicians in the ED and across the healthcare system.

References

[1] Samaras N, Chevalley T, Samaras D, Gold G. Older patients in the emergency department: a review. Annals of emergency medicine. 2010 Sep 1;56(3):261-9.
[2] Esmaeilzadeh P, Mirzaei T. The potential of blockchain technology for health information exchange: experimental study from patients' perspectives. J Med Internet Res. 2019;21(6):e14184.
[3] Rudin RS, Motala A, Goldzweig CL, Shekelle PG. Usage and effect of health information exchange: A systematic review. Ann Intern Med. 2014;161(11):803-11.
[4] Mullins AK, et al. Predictors of Clinician Use of Australia's National Health Information Exchange in the Emergency Department: an Analysis of Log Data. Inter J of Medical Informatics. 2022 Feb 22:104725.

Challenges of Trustable AI and Added-Value on Health
B. Séroussi et al. (Eds.)

doi:10.3233/SHTI220532

DeepUnity Capture: An Application for Digital Photo Documentation

Johanna WEISHAUPT[a,1]
[a] *Dedalus HealthCare*

Abstract. The research on which this poster is based deals with the requirements engineering of an application for medical photo documentation. Against this backdrop, the poster sets out the underlying concepts of medical informatics, the process of medical photo documentation and the standardized procedure of requirements engineering. Using these standards and methodologies, requirements for a mobile photo documentation solution have been elicited, prepared, documented and modeled. As a result of this work, a standardized specification according to ISO/IEC/IEEE 29148, a demonstration model as well as a functional prototype of the application to be designed have been established. From this prototype, an application was developed that is now in routine clinical use and is constantly being refined.

Keywords. Medical photo documentation, Mobile app, Structured documentation, PACS solution

1. Introduction

Regardless of which departments are involved in a hospital, medical data must be available on demand. Particularly in the case of image-generating devices, the aspects of a medical image format must be considered to ensure high image quality. For this reason, the objective of this project has been to conceptualize an application that can be integrated into any existing hospital IT landscape using the healthcare communication standards currently in place. These image recordings are of high medical-legal relevance. It should therefore be possible to create and store them in a standardized manner, meaning within the DICOM format.

2. Methods

Survey:

- Review of the current state and initial requirements elicitation
- Modeling of a standardized process of medical photo documentation
- Structured brainstorming for further functional requirements elicitation

Preparation:

[1] Corresponding Author, Johanna Weishaupt; E-mail: johanna.weishaupt@dedalus.com.

- Categorization and priorization of requirements
- Creation of a Stakeholder and Software Requirements Specification
- Creation of a demonstration model for validation purposes
- Creation of a prototype for further validation

3. Results

The Stakeholder Requirements Specificaton and the Software Requirements Specification provide a complete, standardized outline of functional requirements and thus represent the basis of the development work. The demonstration model is a slide sequence that visualizes both functional and non-functional requirements and thus graphically complements the specification documentation.

The prototype is a development candidate which can be used for validation with stakeholders.

4. Conclusions

Within the defined scope, the relevance of standardized photo documentation for clinical routine has been demonstrated. The app can make a significant contribution to both standardizing photo documentation and streamlining the entire photo documentation process, thereby simplifying it for staff. Nevertheless, at the time of finalizing the poster, the assessment of accompanying circumstances and possible resulting requirements was not yet complete, as this would have exceeded the scope of the thesis. Continuous further development of the app should therefore be considered for successful use in everyday clinical practice.

References

[1] DICOM: Media Storage and File Format for Media Interchange, 2018.
[2] HL7: Semantische Interoperabilität – Terminologien, 2018.
[3] IEEE: Standard Computer Dictionary: A Compilation of IEEE Standard Computer Glossaries. January 1991.
[4] IEEE: Standard Glossary of Software Engineering Terminology. December 1990.
[5] ISO/IEC/IEEE 29148: Systems and software engineering – Life cycle processes – Requirements engineering. December 2011.
[6] Leffingwell D, Widrig D. Managing Software Requirements – A Use Case Approach, Boston: Pearson Education Inc., 2003.
[7] Patrick T. Managing Data, Information and Knowledge. In: Strategic Management of Information Systems in Healthcare 2005. Ed.: Foundation of the American College of Healthcare Executives. Chicago: Health Administration Press, 2005, S. 99-115. Washington DC: AUPHA Press, 2005.
[8] Pedersen S, Hasselbring W. Interoperabilität für Informationssysteme im Gesundheitswesen auf Basis medizinischer Standards, in: Informatik Forschung und Entwicklung, 2004 (18).
[9] Pianykh O. Digital Imaging and Communications in Medicine (DICOM) – A Practical Introduction and Survival Guide, Berlin-Heidelberg: Springer-Verlag, 2008.
[10] Sommerville I. Software Engineering, München: Pearson Deutschland, 9. Auflage, 2012.
[11] Swoboda W. Informationsmanagement im Gesundheitswesen, Konstanz und München: UVK Verlagsgesellschaft mbH, 2017.
[12] Vargas B, Ray P. Interoperability of Hospital Information Systems: A Case Study. In: Proceedings 5th International Workshop on Enterprise Networking and Computing in Healthcare Industry (HealthCom), 2003.

Challenges of Trustable AI and Added-Value on Health
B. Séroussi et al. (Eds.)

doi:10.3233/SHTI220533

Influenza Screening Using Patient-Generated Health Data in Post COVID-19 Pandemic

Hyunwoo CHOO [a], Myeongchan KIM [b], Dohyung LEE [b], and Soo-Yong SHIN [a, 1]
[a] *Department of Digital Health, SAIHST, Sungkyunkwan University*
[b] *Mobile Doctor Co. Ltd, South Korea*

Abstract. It is very important to ensure reliable performance of deep learning model for future dataset for healthcare. This is more pronounced in the case of patient generated health data such as patient reported symptoms, which are not collected in a controlled environment. Since there has been a big difference in influenza incidence since the COVID-19 pandemic, we evaluated whether the deep learning model can maintain sufficiently robust performance against these changes. We have collected 226,655 episodes from 110,893 users since June 2020 and tested the influenza screening model, our model showed 87.02% sensitivity and 0.8670 of AUROC. The results of COVID-19 pandemic are comparable to that of before COVID-19 pandemic.

Keywords. Influenza screening, deep learning, mobile health, Patient generated health data

1. Introduction

Reports of seasonal influenza infection worldwide have declined during the COVID-19 pandemic [1]. This does not mean that the influenza virus has disappeared, but the cases of influenza might have been rarely reported influenza-like illness (ILI) surveillance system because of COVID-19 pandemic. In our previous studies, we demonstrated that it was possible to screen influenza using both surveillance information from patient-generated health data (PGHD) and deep learning model [2]. In this study, we will examine whether a deep learning-based screening tool can still operate effectively even in a situation when the pandemic disrupted ILI surveillance system,

2. Methods

We retrospectively collected data through smartphone application Fever Coach from June 2016 to July 2021 in Korea. The variables are as follows: the subject's age, gender, weight, body temperature measured by the user, the type and dose of the drug taken, symptoms, and diagnosis. The proportion of influenza reports for each week was calculated from the collected Fever Coach data (App surveillance). During the same

[1] Corresponding Author, Department of Digital Health, SAIHST, Sungkyunkwan University, 115, Irwon-ro, Gangnam-gu, Seoul, Republic of Korea; E-mail: sy.shin@skku.edu;

period, weekly ILI surveillance provided by the Korea Centers for Disease Control and Prevention (KCDC surveillance) and daily weather data provided by the Korea Meteorological Administration were collected.

We divided the records into episodes and labeled the episodes as influenza or non-influenza depending on whether an influenza report was in the episode. We assigned corresponding KCDC Surveillance, App Surveillance, and the weather data for each episode by the date of the episode. The episodes before June 2020 were used for training, and the episodes after June 2020 were used for test. In the training set, non-influenza episodes were randomly sampled to equal the number of influenza episodes.

An attention-based deep learning model, called multi-time attention network (mTAN), was used for prediction [3]. Adam was used for optimizer, and cross entropy was used for loss function. 5-fold cross validation was applied during the training. All inputs were zero-mean normalized.

3. Results

From June 2016 to July 2021, a total of 149,234 users entered at least one diagnosis record on the Fever Coach app. We collected 226,655 episodes from the data during the period, and divided 110,893 episodes into the training dataset and other 115,762 episodes into the test dataset. Before the COVID-19 pandemic (June 2020), 15.62% of the training dataset were influenza episodes, which decreased 7.16% in the 2019/2020 season, and even dropped to less than 1% in the 2020/2021 season.

Our model for influenza screening achieved a sensitivity 82.4 % ($\pm$0.16), AUROC 0.898 ($\pm$0.004) in the training set. In the test set, the model achieved a sensitivity 87.02% ($\pm$0.15), and AUROC 0.8670 ($\pm$0.002).

4. Discussion and Conclusion

In a previous study, we showed the possibility of combining deep learning-based influenza screening using patient reported symptoms. There were influenza outbreaks in all three years we previously analyzed. Therefore, the model was neither trained nor tested for years where there was no influenza outbreak. In this study, we found the model maintained in fair performance even though little influenza case has been reported due to the unprecedented COVID-19 pandemic. Further, we showed that deep learning model using PGHD, which are not well-controlled, can have strong performance.

References

[1] Olsen S, Azziz‐Baumgartner E, Budd A, Brammer L, Sullivan S, Pineda R et al. Decreased influenza activity during the COVID‐19 pandemic−United States, Australia, Chile, and South Africa, 2020. American Journal of Transplantation. 2020;20(12):3681-3685.

[2] Choo H, Kim M, Choi J, Shin J, Shin S. Influenza Screening via Deep Learning Using a Combination of Epidemiological and Patient-Generated Health Data: Development and Validation Study. Journal of Medical Internet Research. 2020;22(10):e21369.

[3] Shukla S, Marlin B. Multi-Time Attention Networks for Irregularly Sampled Time Series [Internet]. arXiv.org. 2022 [cited 23 January 2022]. Available from: https://arxiv.org/abs/2101.10318

Challenges of Trustable AI and Added-Value on Health
B. Séroussi et al. (Eds.)

doi:10.3233/SHTI220534

Design of a Patient-Accessible Electronic Health Record System in Mental Health

Stefan HOCHWARTER[1,a,b], Georgia FEHLER[b], Catharina MUENTE[b,d], Yvonne EISENMANN[b], Martin HEINZE[b,c], Charline HIBSCH[b], and Julian SCHWARZ[b,c]
[a] *Norwegian University of Science and Technology, Trondheim, Norway*
[b] *Center for Health Services Research, Brandenburg Medical School Theodor Fontane, Rüdersdorf, Germany*
[c] *Clinic for Psychiatry and Psychotherapy, Brandenburg Medical School Theodor Fontane, Rüdersdorf, Germany*
[d] *Department of Neurology, University Hospital Marburg, Marburg, Germany*

Abstract. Patient accessible Electronic Health Records (PAEHRs) are increasingly implemented internationally. However, studies carried out in the mental health care setting report several practical and ethical challenges when introducing PAEHRs. In this paper we aim to explore the requirements of a PAEHR system in mental health. As part of a participatory design process, we collected qualitative data from service users and staff in a rural mental health day clinic setting, which can be summarized in the following themes: I) Function and way of the documentation; II) Impact on Treatment; III) Concerns about PAEHRs; IV) time of access to PAEHRs; V) Different views on what to share; VI) Access, Data Privacy and Special Features. Our study uncovered the complexity and special requirements and barriers to the design of PAEHR in mental health. While we are in an early stage of our study, we will continue this iterative process and adapt the PAEHR system to the specific needs of the users and domains.

Keywords. OpenNotes, Participation, Mental Health, Psychiatry, Pilot, Co-Design.

1. Introduction

Over the last years, patient access to their electronic patient records has proceeded in many countries and various healthcare domains [1]. Previous research hints at the positive effects of sharing notes with people affected by mental health conditions, yet more research is needed. Specifically, there is little known about how the service users wish to use their notes [2]. Hence, the aim of this study is to explore the requirements on a PAEHR in mental health, which supports both access and interaction with the notes.

2. Method

The study started in June 2021 and is an ongoing research project that follows an explorative and qualitative approach. Data were analyzed inductively and categories emerged from open coding [3]. In this paper, we present preliminary findings. Four focus

[1] Corresponding Author, E-mail: stefan.hochwarter@ntnu.no

group sessions [4] were conducted to co-design the pilot, which subsequently was evaluated in practice. We implemented the feedback between the sessions and discussed the pilot. Nine health professionals (HP) and 14 service users (SU; persons who use mental health services, i.e., patients) participated in these sessions. The study took place in a rural German mental health day clinic.

3. Findings

Both groups (HP and SU) expect that shared documentation will have different effects on the language used in documentation. While HP are concerned that shared documentation can influence their use of language, SU wish respectful but honest documentation. One critical point relates to apparent transparency, i.e., less honest documentation or an incentive to duplicate documentation.

Both HP and SU believe that shared documentation can influence how therapy is carried out. SU report that this can lead to a more trusting relationship between HP and them. On the other side, HP are concerned that this can result in less open conversations. HP and SU both agreed that, besides the possibility of getting more involved in therapeutic processes, there is also a risk for the SU of feeling overwhelmed or overloaded caused by data transparency. Furthermore, HP assumed a lack of comprehensibility of medical and psychological terminology for SU and were therefore concerned about an increased workload. However, both groups also see shared documentation as an opportunity for increased transparency throughout the therapy, which can also uncover discrepancies or mistakes. Finally, they both would welcome a system that allows active participation by the SU, e.g., comments and changes by SU.

4. Discussion and Conclusion

Our findings align with the body of literature on this topic [1]. The findings indicate that before implementing a PAEHR, general reservations and preferences about shared documentation need to be identified and considered. Previous research showed that some concerns would weaken once the users experience PAEHR. At the same time, our findings show the need for a flexible system that fits both the SU's and HP's needs.

The study uncovered requirements and barriers for the design of PAEHR in mental health care. While we are in an early stage of our study, we are impressed by the diversity and high quality of comments and discussions the focus groups provided. Further findings will be published in an upcoming article, and we will iteratively adapt the pilot to the specific needs of the users and domains.

References

[1] DesRoches CM. Healthcare in the new age of transparency. In Seminars in Dialysis 2020 Nov (Vol. 33, No. 6, pp. 533-538).
[2] Schwarz J, Bärkås A, Blease C, Collins L, Hägglund M, Markham S, Hochwarter S. Sharing Clinical Notes and Electronic Health Records With People Affected by Mental Health Conditions: Scoping Review. JMIR mental health. 2021 Dec 14;8(12):e34170.
[3] Flick U, editor. The SAGE handbook of qualitative data analysis. Sage; 2013 Dec 18.
[4] Krueger RA. Focus groups: A practical guide for applied research. Sage publications; 2014 Aug 14.

Challenges of Trustable AI and Added-Value on Health
B. Séroussi et al. (Eds.)

doi:10.3233/SHTI220535

Discovery of Biomedical Databases Through Semantic Questioning

Arnaldo PEREIRA[a,1], João Rafael ALMEIDA[b], Rui Pedro LOPES[c], Alejandro PAZOS[b], and José Luís OLIVEIRA[a]
[a] *DETI/IEETA, University of Aveiro, Portugal*
[b] *Department of Computation, University of A Coruña, Spain*
[c] *CeDRI, Polytechnic Institute of Bragança, Portugal*

Abstract. Many clinical studies are greatly dependent on an efficient identification of relevant datasets. This selection can be performed in existing health data catalogues, by searching for available metadata. The search process can be optimised through questioning-answering interfaces, to help researchers explore the available data present. However, when searching the distinct catalogues the lack of metadata harmonisation imposes a few bottlenecks. This paper presents a methodology to allow semantic search over several biomedical database catalogues, by extracting the information using a shared domain knowledge. The resulting pipeline allows the converted data to be published as FAIR endpoints, and it provides an end-user interface that accepts natural language questions.

Keywords. Clinical Studies, Health Data, Ontology, Semantic Questioning

1. Introduction

Over the last years, the secondary use of data from medical and biomedical practice gained attention from the health community, leading to new strategies aiming to enable access to clinical data from distributed databases without losing patient data privacy [1]. These strategies help researchers identify datasets of interest to design and conduct a multi-institutional study. One of the goals of the European Medical Information Framework (EMIF) (http://www.emif.eu) project was to improve the access of researchers to patient-level data from distinct health databases across Europe. In this project, EMIF Catalogue was developed to provide metadata information about each database to overview the database content without exposing the data.

Identifying databases of interest is one of the challenges researchers face in this platform. The main issue is the lack of a strategy to correlate concepts that are similar in distinct databases, for instance, procedures that can have different designations due to separate institutional policies. This work proposes a strategy to extract biomedical data through an ontology that guides the data conversion into a semantic format. To facilitate access to this semantic repository, we propose an interface that receives queries written in natural language.

[1] Corresponding author: E-mail: arnaldop@ua.pt

2. Methods and Discussion

The proposed methodology aims to convert questions written in the English language into formal queries based on question-answering templates. There is a clear benefit in simplifying the access to semantic data and saving time to end-users. For instance, it enables the evaluation of a research study's feasibility before spending time in the definition of a study protocol for a topic that may not be viable.

In this use case, we want to improve medical researchers' user experience when navigating the database catalogue by retrieving databases with free-text questions. This feature has some limitations due to the data scope. However, it can expand the usual match-based features available on this platform. This methodology was implemented as a plugin in the EMIF Catalogue, which requires the manual mapping of the catalogue concepts to the ontology concepts. This initial effort creates a new description layer for the catalogue entities, as well as, it also provides additional knowledge to each concept. For instance, with ontology is possible to have a relation between the entities, while in the original version, there is no relation.

3. Results and Conclusions

The proposed system was integrated into the EMIF Catalogue platform as a plugin which was validated in the study of Alzheimer's disease containing the metadata of 130 cohorts, of which 62 are public available. Each dataset is characterised by 472 entities. The searching features in this platform are provided through exact matching or through the use of boolean operation to build a nested query. With the inclusion of the proposed methodology, the search capabilities were expanded, allowing less restricted searches.

In a previous work, we used a recommender system to enable researchers to discover cohorts of interest [2]. Although these systems provide good results, they are not suitable for verifying studies' feasibility. The use of semantic queries over biomedical databases enables the exploration of data through a higher level of representation. However, to take benefit of this technology, it is required some technical background and deep knowledge of the ontology used in each scenario.

Acknowledgments

This work has received support from the EU/EFPIA Innovative Medicines Initiative 2 Joint Undertaking under grant agreement No 806968. AP and JRA are funded by the FCT - Foundation for Science and Technology (national funds) under the grants PD/BD/142877/2018 and SFRH/BD/147837/2019 respectively.

References

[1] Almeida JR, Fajarda O, et al. Strategies to access patient clinical data from distributed databases. In HEALTHINF, pages 466–473. SciTePress, 2019.

[2] Almeida JR, Monteiro E, et al. A recommender system to help discovering cohorts in rare diseases. In 2020 IEEE 33rd International Symposium on Computer-Based Medical Systems (CBMS), pages 25–28. IEEE, 2020.

Section V

Citizen Health, Public Health and Epidemiology Informatics

This page intentionally left blank

Challenges of Trustable AI and Added-Value on Health
B. Séroussi et al. (Eds.)

doi:10.3233/SHTI220537

Over 89% Adoption Rate of the Nationwide Online Patient Portal in Finland

Vesa JORMANAINEN[a,b1]
[a] *Finnish Institute for Health and Welfare, Helsinki, Finland*
[b] *University of Helsinki, Department of Public Health, Helsinki, Finland*

Abstract. Among 1,650 persons in an internet panel survey in October 2020 in Finland, over 89% reported use of the nationwide My Kanta online patient portal. Only 1.5% of the respondents did not know the service. Compared with non-users, among My Kanta users there were more females, less living in countryside, household net income was higher, and more reported independent use of online services. My Kanta use increased by poorly self-rated health status, increasing number of reported prescribed medicines, long-term diseases and physician visits during the six previous months.

Keywords. Patient portals, health information systems, My Kanta, Finland

1. Introduction

A patient portal is a service that enables patient to access their health information online [1,2]. Patients are satisfied with their access to their own health data [3–7]. Shared nationwide patient portals have been introduced in many countries [8–17], but only few studies have reported the use of nationwide patient portals [10,13–14,18].

Assessment of adoption rates is essential in understanding effects of patient portals on decision-making, care processes and health outcomes [19]. An overall 52% adoption rate has been observed, but it was 71% [95% CI: 65–79%] in controlled and 23% [95% CI: 13–33%] in real-world experiments.

The highest use of electronic prescription and consultation was in Finland among 28 European Union Member States in 2016 [20]. Finland has introduced since May 2010 national, centralized, shared, integrated and interoperable electronic data system services (Kanta Services) in phases for citizens, healthcare, social welfare and pharmacy service providers [18,21]. My Kanta is a nationwide online service allowing citizen service users to view their information about electronic prescriptions and health data via an internet web page. My Kanta use has increased during the 10 years from 16% in 2014 to 64% in 2020–2021 [15,22–23] and to as high as 83% among community pharmacy customers in 2019 [18,24].

In this study, we aimed to assess adoption rates (use proportions) of the nationwide My Kanta online patient portal in Finland by user and non-user characteristics.

[1] Corresponding Author: MD Vesa Jormanainen, Finnish Institute for Health and Welfare (THL), P.O. Box 30, 00271 Helsinki, Finland; E-mail: vesa.jormanainen@thl.fi.

2. Methods

A cross-sectional web-based questionnaire survey took place in the Taloustutkimus Oy's internet panel (approximately 40,000 members) in October 15–31, 2020 among Finnish persons aged 18–79-years of age, who were resident in Finland [25]. A random stratified sample of 9,466 panel members was constructed and invited to the questionnaire survey. The quantitative response goal (1,500) took place in October 31, 2020.

Use of My Kanta was assessed by asking the question "Have you used My Kanta patient portal to access your electronic prescriptions and/or other health data?". The response categories were: i) Yes, electronic prescription data, ii) Yes, other health data, iii) Yes, both electronic prescription and other health data, iv) No, and v) I do not know My Kanta. In data processing, the response categories i–iii were classified into "Yes" (users) and iv–v into "No" (non-users).

3. Results

Totally 1,650 persons responded to the internet panel survey. Of the respondents, 89.5% used My Kanta, and 10% of the users accessed their prescription data only, 10% their other health data only and 81% both. Eleven percent did not use My Kanta and 1.5% did not know the service at all.

There were more females among patient portal users (56%) than among the non-users (31%), and more non-users (23%) than users (12%) reported living in countryside. Household monthly median net income was higher among users than the non-users. In addition, 97% of the users (86% of non-users) used online services independently without any help.

3.1. Use of the nationwide My Kanta patient portal

Females (93%) reported higher use than males (82%). The use proportion was highest in the oldest (65–79-year-olds) (90%) and in the youngest (18–34-year-olds) (89%) age groups. My Kanta use was lowest in education level elementary school (80%) and highest in university (91%) education level. Pensioners (90%) reported the highest use as well as those who were on maternity leave or students (89%).

3.2. Use of My Kanta patient portal by respondents' health-related characteristics

My Kanta use increased by poor self-rated health status: rather good or good (87%), neither good nor poor (89%), and rather poor or poor (95%). Responsible caregivers of family members or relatives reported higher use (98%) than the other groups (88%). My Kanta use also increased by increasing number of prescribed medicines: no prescription medicines (72%), 1–2 (89%), 3–4 (94%), 5–9 (96%), and 10 or more (100.0%). In addition, number of reported physician-diagnosed long-term diseases showed an increasing trend: no long-term disease (79%), one (87%), two (93%), three (95%), 4–5 (95%), and 6+ (100%).

The reported use of My Kanta was highest (100%) among the respondents who had a physician-diagnosed epilepsy (prevalence 1.2%), lowest (90%) among those who reported a skin disease (prevalence 8.7%) and 79% among those who did not report any

long-term diseases (prevalence 31%) (Figure 1). In addition, use of My Kanta increased by reported physician visits during the previous six months: none (81%), 1–2 (89%), 3–5 (94%), 6–10 (95%) and more than 10 (95%).

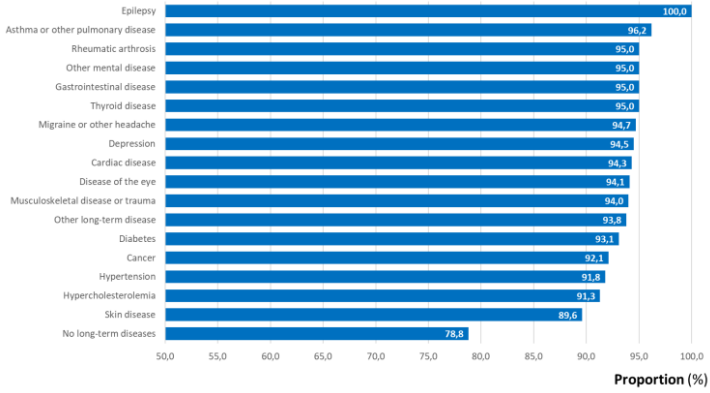


Figure 1. The proportion (%) of the nationwide My Kanta patient portal by self-reported physician-diagnosed long-term diseases among 1,650 respondents in Taloustutkimus Oy's internet panel survey among Finnish persons aged 18–79-years of age and who were resident in Finland inOctober 15–31, 2020 in Finland.

4. Discussion and Conclusion

Among 1,650 persons in an internet panel survey in October 2020 in Finland, over 89% reported use of the nationwide My Kanta online patient portal, some 10% of the users accessed their prescription data only, and 10% accessed their other health data only and over 80% both. Only 1.5% of the respondents did not know the patient portal service.

My Kanta use has increased significantly in less than 10 years [18,22–24]. In this survey My Kanta use was high, and nationwide patient portal adoption rates in Finland are in general higher than in real-world or controlled experiments [19].

Among My Kanta users compared with non-users, there were more females, less were living in countryside, household net income was higher, and more reported independent use of online services. These results may indicate better health literacy, access to internet and online service skills among the internet panel members in general. My Kanta use increased by poorly self-rated health status, increasing number of reported prescribed medicines, long-term diseases and physician visits during the six previous months. These results are likely associated to the use of patient portal as an essential module in cure and care processes. It also shows that there exist new and continuously relevant and important content for the users in the nationwide patient portal.

This study results suggest that it is possible to introduce a nationwide patient portal that is available for all potential users, and system availability leads to ongoing and increasing portal use [26]. Next studies should investigate if increasing patient portal use would be associated with observable changes in clinical and health behavior that further likely would result in improvements in patient outcomes.

References

[1] Osborn VY, Mayberry LS, Wallston KA, et al. Understanding patient portal use: implications for medication management. J Med Internet Res 2013;15(7):e133.
[2] Essén A, Scandurra I, Gerrits R, et al. Patient access to electronic health records: differences across ten countries. Health Policy Technol 2018;7(1):44–56.
[3] Delbanco T, Walker J, Bell S, et al. Inviting patients to read their doctors' notes: a quasi-experimental study and a look ahead. Ann Intern Med 2012;157(7):461–470.
[4] Turvey C, Klein D, Fix G, et al. Blue Button use by patients to access and share health record information using the Department of Veterans Affairs' online patient portal. J Am Med Inform Assoc 2014;21(4):657–663.
[5] Esch T, Mejilla R, Anselmo M, et al. Engaging patients through open notes: an evaluation using mixed methods. BMJ Open 2016;6(1).
[6] Walker J, Leveille S, Bell S, et al. Open notes after 7 years: patient experiences with ongoing access to their clinicians' outpatient visit notes. J Med Internet Res 2019;21(5):e13876.
[7] Wass S, Vimarlund V, Ros A. Exploring patients' perceptions of accessing electronic health records: innovation in healthcare. Health Inform J 2019;25(1):205–215.
[8] de Lusignan S, Ross P, Shiffrin M, et al. A comparison of approaches to providing patients access to summary care records across old and new Europe: an exploration of facilitators and barriers to implementation. Stud Health Technol Inform 2013;192:397–401.
[9] Hardardottir GA, Thoroddsen Å. National eHealth implementation: country experience. Stud Health Technol Inform 2016;225:168–172.
[10] Lämsä E, Timonen J, Mäntyselkä P, et al. Pharmacy customers' experiences with the national online services for viewing electronic prescriptions in Finland. Int J Med Inform 2017;97:221–228.
[11] Nohr C, Parv L, Kink P, et al. Nationwide citizen access to their health data: analysing and comparing experiences in Denmark, Estonia and Australia. BMC Health Serv Res 2017;17(1):534.
[12] Oderkirk J. Readiness of electronic health record systems to contribute to national health information and research. OECD Health Working Paper 99. Paris: OECD, 2017.
[13] Moll J, Rexhepi H, Cajander Å, et al. Patients' experiences of accessing their electronic health records: national patient survey in Sweden. J Med Internet Res 2018;20(11):e278.
[14] Holt KA, Karnoe A, Overgaard D, et al. Differences in the level of electronic health literacy between users and nonusers of digital health services: an exploratory survey of a group of medical outpatients. Interact J Med Res 2019;8(2):e8423.
[15] Jormanainen V, Parhiala K, Niemi A, et al. Half of the Finnish population accessed their own data: comprehensive access to personal health information online is a cornerstone of digital revolution in Finnish health and social care. Finnish J eHealth eWelfare 2019;11(4):298–310.
[16] Zanaboni P, Kummervold PE, Sorensen T, et al. Patient use and experience with online access to electronic health records in Norway: results from an online survey. J Med Internet Res 2020;22(2):e16144.
[17] Carini E, Villani L, Pezzullo A, et al. The impact of digital patient portals on health outcomes, system efficiency, and patient attitudes: updated systematic literature review. J Med Internet Res 2021;23(9):e26189.
[18] Sääskilahti M, Ahonen R, Timonen J. Pharmacy customers' experiences of use, usability, and satisfaction of a nationwide patient portal: survey study. J Med Internet Res 2021;23(7):e25368.
[19] Fraccaro P, Vigo M, Balatsoukas P, et al. Patient portal adoption rate: a systematic review and meta-analysis. Stud Health Technol Inform 2017;245:79–83.
[20] Dubois H. Healthcare. In: Eurofound. Quality of health and care services in the EU. Publications Office of the European Union, 2017.
[21] Jormanainen V. Large-scale implementation and adoption of the Finnish national Kanta services in 2010–2017: a prospective, longitudinal, indicator-based study. Finnish J eHealth eWelfare 2018;10(4):381–395.
[22] Hyppönen H, Hyry J, Valta K, et al. Electronic services in the social welfare and health care sector: citizens' experiences and development needs. Report 33/2014. Helsinki (Finland): National Institute for Health and Welfare, 2014.
[23] Kyytsönen M, Aalto A-M, Vehko T. Social and health care online service use in 2020–2021: experiences of the population. Report 7/2021. Helsinki (Finland): Finnish Institute for Health and Welfare (THL), 2021.
[24] Sääskilahti M, Aarnio E, Lämsä E, et al. Use and non-use of a nationwide patient portal: a survey among pharmacy customers. J Pharmaceutical Health Serv Res 2020;11(4):335–342.

[25] Saastamoinen L (Ed.), Airaksinen M, Dimitrow, et al. Creating price competition in the pharmaceutical market and the public's expectations of pharmacies. Publications of the Governments' analysis, assessment and research activities 2021:32. Helsinki (Finland): Prime Minister's Office, 2021.
[26] Price M, Lau F. The clinical adoption meta-model: a temporal meta-model describing the clinical adoption of health information system. BMC Med Inform Decision Mak 2014;14:43.

Challenges of Trustable AI and Added-Value on Health
B. Séroussi et al. (Eds.)

doi:10.3233/SHTI220538

Understanding Public Priorities and Perceptions of the Use of Linked Healthcare Data in South East England

Elizabeth FORD [a,1], Kathryn STANLEY [a] Melanie REES-ROBERTS[b], Sarah GILES[c], Katie GODDARD[a], Jo ARMES[d]
[a] *Brighton and Sussex Medical School, Brighton, UK*
[b] *University of Kent, Canterbury, UK*
[c] *Public Advisor, Kent, UK*
[d] *University of Surrey, Guildford, UK*

Abstract. The counties of Kent, Surrey and Sussex (KSS) in South East England are creating anonymized, linked databases of healthcare records for audit, service planning and research for the first time. We consulted with 79 citizens from KSS in 5 deliberative focus groups, asking about perceived benefits and concerns regarding these new data assets. Participants hoped the linked datasets could be used for joining up care and information, improving efficiency, and improving healthcare provision, but were concerned about missing and inaccurate data, data breaches and hacking, use of data by profit-making organisations, and stigma and discrimination. Findings will be used to underpin governance and engagement strategies for integrated datasets in KSS.

Keywords. Linked data, patient data, public engagement, data governance, privacy

1. Introduction

In England, the government has given national funding to each region to set up digital transformation programmes in the National Health Service (NHS). Each area, generally at a county level, now has an integrated care system (ICS), bringing together previously separated healthcare providers under one umbrella organization [1]. The transformation includes plans to create an extracted, anonymized and linked patient record database for audit, service planning, commissioning and research. This is the first time that the counties of Kent, Sussex, and Surrey (KSS) in South East England have had such health data infrastructure. The NIHR-funded "Unlocking Data to Inform Public Health Policy and Practice" project (NIHR133761) has brought together universities, local government, and NHS to use these emerging datasets for public health planning and policy.

One of the key issues to consider when linking and curating data from patients' health records is how the public view this activity. We saw an opportunity to engage with citizens of Kent, Surrey and Sussex to create data governance structures from the start which were inclusive, transparent, and which have citizen support. There was very little

[1] Corresponding Author, Dr E Ford, Department of Primary Care and Public Health, Brighton and Sussex Medical School, Watson Building, Village Way, Falmer, BN1 9RH e.m.ford@bsms.ac.uk.

local evidence of public awareness, or "reasonable expectations" of patients and public around population data linkage for analytics. A publicly-supported and informed data governance and access structure in KSS would ensure uses of patient data are appropriate, ethical, and can be prioritized against public benefit and need. According to social licence theory, the public expect that organisations who, for example, hold and use the public's data, will go beyond the requirements of formal regulation, and adhere to voluntary codes of trustworthy behaviour and transparency [2]. Where the public are satisfied that the motivations and competency of the organisation are trustworthy, they may confer a "social licence" to operate. Additionally, the public consider core research ethics principles when weighing up approval of data-sharing and linkage schemes [3-5].

In this study we aimed to engage with citizens of Kent, Sussex and Surrey, to understand views around linked health dataset use, particularly, what benefits they see will accrue for the public from these assets and what concerns they have about their use.

2. Methods

2.1. Study Design

This project was approved by Brighton and Sussex Medical School Research Governance & Ethics Committee (ref: ER/BSMS2730/7). We conducted deliberative discussion focus groups according to the methodology of Rothwell et al [6]. Deliberative research is an approach for gathering wider views about health issues when there are many complex issues to weigh against each other [4]. Deliberative discussion focus groups include a range of informative presentations, interspersed with facilitated plenary and breakout discussions.

2.2. Recruitment and Procedure

We advertised our study through a large range of health, social care and support organisations in Kent, Surrey and Sussex via an online flyer, giving a link to a website at which potential participants could express an interest. This was followed by sending participants a full information sheet and consent form. Once consented, participants were booked into a focus group and were sent further online reading about issues to be discussed. Participants were sent a £50 voucher after completing the focus group.

Focus groups were conducted remotely using Zoom video-conferencing software and were recorded. Each group consisted of two sessions one week apart totalling 3 hours; at the beginning, participants were informed both benefits and risks would be explored. All participants completed an online pre-group demographics questionnaire and post-group feedback questionnaire; both included a question on willingness to share their medical records to ensure a diverse range of views were represented.

2.3. Focus group content and discussion questions.

Session 1: What integrated datasets of health and social care data could be used for in public health (with case studies on joining up care for diabetes and multi-morbidity).

Q1: What could be the benefits of using these data for research and commissioning for the public?

Q2: What can be achieved with these datasets that otherwise would not happen?
Q3: What are the most important types of projects that should be done with these data? What do you feel is less important or should not be done?

Session 2: Governance and privacy issues, concerns and solutions for linked data.
Q1: Do you have any concerns or suggestions around the current framework for protecting health and social care data in integrated datasets?
Q2: What are your views on data being linked together, particularly health and non-health data (e.g. housing benefits, crime reporting, or educational attainment data)? Do you have any particular worries about this?
Q3: Who should have access to these datasets for analysis, and who should we be cautious about giving access to? What vetting procedure should they go through?

2.4. Data Analysis

Questionnaire responses were analysed descriptively. Zoom discussion groups were audio-recorded and transcribed verbatim. Written transcripts were uploaded to NVivo (version 1.4.1) for thematic analysis according to the 6-steps of Braun and Clarke [7]. Firstly, transcripts were read repeatedly and initial ideas noted down, secondly interesting features of the data were coded. Codes were then collated into potential themes and a thematic map was produced. Then authors reviewed the specifics of each theme against the study aims, the overall story, and the thematic map, and selected extracts which best represented each theme.

3. Results

Of 152 people who expressed an interest to take part, 79 completed one of 5 focus groups.

Table 1. Demographics of the sample (ND: Not disclosed)

Characteristic	N (Total = 79)
Gender	Female: 52 Male 26 Non-Binary: 1
Ethnicity	White 63, Asian 10, Black 5, Mixed 1
Educational Attainment	School leaver 23; University education 44; ND: 12
County of Residence	Kent 11, Sussex 49 Surrey 19
Disability Status	Yes: 17 No: 47; ND: 15
Willingness to share medical data (pre)	Willing 62 Unwilling 15 Don't know 2
Willingness to share medical data (post)	Willing 65 Unwilling 7 Don't know 7

The key themes expressed by participants were around (1) the potential **benefits** of using integrated data, and (2) the potential **risks or concerns**.

3.1. Benefits

The benefits expected from linking data for service planning and research included **joining-up care and information, improving efficiency, and improving healthcare provision.**

Many participants referred to prior experiences of fragmented care provision which focused on single issues or where different parts of the health system were not aware of what other providers were doing. Participants hoped that the new linked datasets would help planners to commission more joined-up services.

"All this dataset analysis and helping to link them together will enable each department to do their job better. All of them are wasting so much time in doing the same thing again or doing the wrong thing because they haven't got the information. This will make everything so much more efficient for everybody."

Participants also hoped that better information flows would reduce the need to repeat their stories to multiple clinicians, wasting time and effort, and risking key clinical information being lost or missed:

"The efficiency for the NHS and all the other players will be in the non-duplication of all these visits ...I explained a problem I have to my GP; I then go along to the hospital and then I explain it to a nurse, I then go and see a consultant and rather strangely he didn't know anything about it, 'can you tell me what your problem is?'"

Participants also hoped that the system overview gained by analysing linked health service data would lead to more efficient use of resources, especially targeting care to the appropriate places or the appropriate people:

"If they're wasting less because they're able to stop duplicating things and stuff like that, then everyone's going to get better care overall because the funding won't be used inappropriately and wasted."

Lastly, participants hoped that the joining-up and efficient use of resources would result in improved healthcare, and improved health in KSS, including mental health.

"I think with physical disabilities, with other areas, ...sometimes the doctor will say… we need evidence-based therapies. The only way we're going to have this evidence base is if we start linking data and collecting data in an appropriate way so it can be properly analysed and then used… The outcomes will be so, so much better than what we have today."

3.2. Perceived risks and concerns

Patients noted a number of concerns about the uses of linked datasets, which centred on **missing and inaccurate data, data breaches and hacking, use of data by profit-making organisations, stigma and discrimination**. Participants were concerned that the data held in linked healthcare records would not be "clean" or accurate, and therefore, any planning or research conducted using the data, would also not be reliable.

"How do we ensure the cleanliness of the data and make sure that the data is accurate enough to build the right picture?"

Furthermore, participants were worried that the data could be hacked, or that anonymization was not a secure way to protect privacy. They were additionally concerned with who might get access to the data and suggested users of data should leave an audit trail *"a bit like you swipe a key card in a building."*

"I don't trust personally that all of this sharing of information isn't going to get hacked. I don't trust the fact that it's all going to be anonymised, there's human error that's going to come in here."

Part of these concerns focused on the issue that participants did not want their data to be used for a profit motive. They felt that research paid for by drug or insurance companies would be less transparent, and less in the public interest.

"Researchers are funded by people, organisations. Who they are funded by and what is their ultimate goal? For example, drug companies fund a huge amount of research but they will not publish those research results if it's not advantageous to them."

Lastly, participants were worried about stigma; both in terms of feeling uncomfortable with data on sensitive issues being shared for planning and research, and also worrying about stigma or discrimination which could arise from misuse of the data.

> *"I think people may feel there's some stigmas to things like [welfare] benefits and I think the stigmas and mental issues, and also sexual issues, is why the health clinics are always stringent about being confidential because people find these issues really difficult."*

4. Discussion and Conclusion

We identified core issues of both support and concern for the linking of anonymised NHS patient records in South East England to facilitate healthcare audit, planning and research in the region. These findings are similar to concerns expressed by patients in previous literature [3,4,8], although previous research has not reported on the public identifying increased efficiency of health services as a benefit of data-linkage. Concerns about poor data quality were also not prominent themes in prior research. We identified that citizens of KSS are keen to see data used to improve healthcare delivery, efficiency and connection of services, but also seek reassurance that data is appropriately safeguarded, and will only be used for the public good. We note limitations of our study: starting with presenting on and discussing the benefits of the datasets may have primed the participants to feel more positive about uses of data, although we did inform participants at the start that we would discuss both benefits and risks. Our findings may not generalize outside KSS, although congruence with national research is largely good.

Findings will enable the ICSs in KSS to create publicly supported governance, engagement and communication strategies, securing a social license for the use of linked health data in the region.

References

[1] The King's Fund. Integrated care systems explained: making sense of systems, places and neighbourhoods. 2021 Available from: https://www.kingsfund.org.uk/publications/integrated-care-systems-explained.

[2] Gunningham N, Kagan RA, Thornton D. Social License and Environmental Protection: Why Businesses Go Beyond Compliance. Law & Social Inquiry. 2004;29(2):307-41.

[3] Stockdale J, Cassell J, Ford E. "Giving something back": A systematic review and ethical enquiry into public views on the use of patient data for research in the United Kingdom and the Republic of Ireland. Wellcome Open Res. 2019;3:6-.

[4] Ford E, Oswald M, Hassan L, Bozentko K, Nenadic G, Cassell J. Should free-text data in electronic medical records be shared for research? A citizens' jury study in the UK. Journal of Medical Ethics. 2020;46(6):367-77.

[5] Ford E, Boyd A, Bowles JKF, Havard A, Aldridge RW, Curcin V, et al. Our data, our society, our health: A vision for inclusive and transparent health data science in the United Kingdom and beyond. Learning Health Systems. 2019;3(3):e10191.

[6] Rothwell E, Anderson R, Botkin JR. Deliberative Discussion Focus Groups. Qual Health Res. 2016;26(6):734-40.

[7] Braun V & Clarke V. Using thematic analysis in psychology. Qualitative Research in Psychology, 2006:3(2), 77-101.

[8] Aitken M, de St Jorre J, Pagliari C, Jepson R, Cunningham-Burley S. Public responses to the sharing and linkage of health data for research purposes: a systematic review and thematic synthesis of qualitative studies. BMC medical ethics. 2016 Dec;17(1):1-24.

Challenges of Trustable AI and Added-Value on Health
B. Séroussi et al. (Eds.)

doi:10.3233/SHTI220539

Why Do People Use Online Lab Results and What Do They Look For: A Qualitative Study

Helen MONKMAN[a,1], Janessa GRIFFITH[b], Leah MACDONALD[a], Amanda L. JOSEPH[a] and Blake LESSELROTH [c]
[a] *School of Health Information Sciences, University of Victoria, Canada*
[b] *Health Information Management, Douglas College, Canada*
[c] *University of Oklahoma-Tulsa, School of Community Medicine, USA*

Abstract. Laboratory (lab) test results are increasingly available online for patient review. However, there is a dearth of research with respect to users' information needs, goals, and information processing strategies. In this exploratory qualitative study, we interviewed a sample of (N = 25) online lab results users to understand their objectives and search targets. We transcribed their responses and used affinity diagramming to identify themes in their responses. Our analysis identified six reasons why people look at their online lab results (i.e., health status, reassurance, health education, speed, self-management, and patient safety) and two themes about what people look for (i.e., abnormal and normal values, trends). Knowing what drives users and what information they are looking for can inform the design of online lab reporting, improve usefulness, and better satisfy user needs.

Keywords. Laboratory test results, patient portals, consumer health information, information needs

1. Introduction

By 2015, all Europeans should have had access to their own medical records [1]. Years later, direct-to-consumer digital health technologies have shown to be an important part of any health promotion strategy, by increasing patient knowledge, engagement, and empowerment. Patient (i.e., citizen, health consumer, layperson) empowerment has been shown to improve health outcomes [2]. Electronic medical records and related technologies with patient-facing portals can inform, empower, and engage citizens by providing access to diagnostic results (e.g., labs), notes, and other health data. Preliminary evidence suggests that access to online lab results increases patient engagement (e.g., self-management, trend monitoring [3], preparing for consultations with health care providers [4]) and also offers reassurance about one's health status [4]. Although patient access to online lab results is now commonplace in many countries, research focusing upon their use and usability has only recently gained traction. For example, some studies have explored patient search results and methods to distinguish between normal and abnormal values (e.g., [5]), but users' motivations for using these

[1] Helen Monkman, School of Health Information Science, University of Victoria, Victoria, British Columbia, Canada; E-mail: monkman@uvic.ca

portals and what information they seek is understudied. User goals and needs should be a driving force behind the design of patient-facing systems. Some researchers have incidentally identified reasons why people review, or should consider reviewing, online lab results (e.g., speed of receiving the result, identifying overlooked results that warrant attention [6]),. However, researchers have not expressly studied this phenomenon and we hypothesize there are additional motivations and priorities driving user search behaviors. We sought to address this gap by examining user goals to determine 1) why people look at their online lab results, and 2) what information they seek. Therefore, we asked these questions to people accessing their online lab results and analyzed their qualitative responses to identify common themes.

2. Methods

The University of Victoria's Human Research Ethics Board approved this study. We recruited participants using a provincial platform for health research. First, participants completed a questionnaire with demographic questions (e.g., age, language, education) and questions about their online lab results experiences. We then invited them to be interviewed using ZOOM™ videoconferencing software [7]. The interviews explored user goals, context of use, and performance using different formats of online lab results displays. The focus of this study is on what motivates the use of online lab results. Therefore, we examined participants' responses to the following questions:

1. *Why* do you look at your online lab results?
2. *What* do you look for?

We recorded and transcribed the interviews and used affinity diagramming to analyze responses and identify themes [8,9]. For this phase of the research, we created one slide for each question in PowerPoint® and then created an individual note for each participant response to each of the two questions. The authors met during a videoconference to review participant responses. We examined each response individually, and then discussed to reach a consensus whether each response represented a novel or existing theme(s). We moved a novel response into its own new group, and we moved responses belonging to an existing theme into its respective group. We assigned names for each theme and modified the structure as necessary throughout the sorting process. We duplicated responses illustrating multiple themes and put a copy of the response in each corresponding thematic group. Participants could only be represented once per theme, but a response could be counted in multiple themes. Although we initially analyzed the responses for each question separately, we found overlap in participant responses (e.g., when we asked what participants were searching for, they would also describe why they were searching). Therefore, in some circumstances we coded responses to the other question rather than the one that was posed. Finally, we tallied the number of responses for each theme.

3. Results

This study was part of a larger study and some results have been accepted for publication [10]. We interviewed 25 participants, representing age categories from 18-74 years old. All primarily spoke English at home, twenty participants were born in Canada and 23

had earned a post-secondary degree. Eighteen participants in this sample had at least one chronic illness and eight participants reported five or more chronic illnesses. All subjects had at least two years' experience using an online lab results portal and nearly all of them (i.e., 24) accessed their online lab results several times a year or more.

3.1. Why do people look at their online lab results and what do they look for?

For brevity, we will only describe themes discussed by five or more participants. We identified six themes indicating *why* people look at their online lab results: health status, reassurance, health education, speed (i.e., timeliness), self-management, and patient safety (**Figure 1**). **Table 1** lists each theme with a supporting quote. Participants wanted to know about their current health status, their lab trends, and if they were responding to treatments (Health Status). Participants said viewing their normal results promptly reassured them (Reassurance). Participants wanted to know more about the tests conducted and the meaning of their results (Health Education). Participants also appreciated how rapidly their results were available for review and that they did not have to wait to get their results from their health care provider (Speed). Another theme identified was that participants used the online lab portal to manage their own conditions (Self-Management). Finally, by reviewing their lab results, participants could detect issues in their results and values that may be overlooked by their health care providers (Patient Safety). In one instance, a participant sought immediate medical assistance due to a lab value (see **Table 1**).

We identified two themes associated with the second question (i.e., *What* do you look for?). Participants commonly reported looking for abnormal findings, normal values, and trends. Participants typically described looking for abnormal values first and some would subsequently look at the normal values in more detail. Several participants also examined trends or the pattern of their results over time.

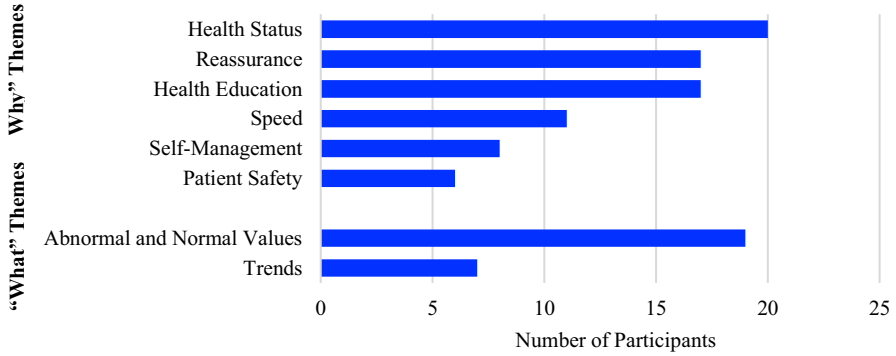


Figure 1. Number of participants whose responses represented each theme. "Why" themes depict reasons why people look at their online results and "what" themes indicate what people look for in them.

Table 1. Themes, descriptions, and quotes of why people look at online lab results and what they look for.

	Theme	Description	Illustrating Quote
Why do people look at	Health status	People want to know their lab values and health status. For example, observing trends in lab values and treatment response.	P19: Well, I'm naturally curious and controlling and like to have and gather information…I also do a lot of comparison…I find the data really useful to tell me whether or not strategies, health strategies, I'm using are hitting the marks or not.

	Reassurance	People want to know whether or not there is cause for concern	P14: And then also like peace of mind, it's pretty easy to interpret like basic lab results. Like if you got tested for strep throat, it's easy to go on and see "Oh, no, I don't have it. I'm fine."
	Health education	People want to know more about the tests performed and the clinical implications.	P6: I think it's just important to review them and try to understand them because it's part of sort of maintaining my own health.
	Speed	People want access to results as soon as possible.	P7: Oftentimes, I'll get that information before my physician gets back to me. So, I like to be informed as quickly as possible.
	Self-management	People can make adjustments to their treatment plans based upon their results.	P4: For my monthly tests, I look for being out of the range that the results should be. And one of them I can start to change my medication that's related to that immediately.
	Patient safety	People can ensure lab results are not overlooked and possibly act sooner on abnormal test results.	P25: One of the results had said that I was in heart failure, so I went to the hospital immediately.
What do they look for?	Abnormal and Normal Values	People look for values within and outside of the reference ranges.	P2: One of the first things I look at is like, are they in the range of normal…mostly, what's in the normal [range], what's abnormal?
	Trends	People look for how their values change over time.	P19: I keep a file folder, and I have things highlighted, and I track them over time. So like, long range, where was I at a year ago? Where was I three months ago, three years ago?

4. Discussion

This study contributes to published data explaining why people review online lab results including reviewing health status [4], quick access to results [6], ensuring results are not overlooked [6], facilitating self-management [3], and monitoring trends [3]. Perhaps most importantly, this study showed that patients view online lab results as a safety mechanism to identify abnormal results and alert them when to seek medical attention. Unfortunately, clinics often fail to communicate abnormal results to patients [11]. Furthermore, clinically important results may not be flagged and could be missed by both patients [10] and providers. We identified new themes including people seeking reassurance and a curiosity to learn about their own health. Participants' informational needs often required additional online research, which has also been described by other studies (e.g., [12]).

Participants typically looked for abnormal and normal results and lab trends. Given the small size and the characteristics (e.g., highly educated, chronic illness) of this sample, some findings may not be generalizable to the population. For example, most participants had one or more chronic illness, increasing their likelihood of tests conducted and the need for monitoring their values over time.

The findings from this study can inform future lab results portals to ensure they are designed to best meet user requirements. For example, future designs should provide notifications to users when their results are ready or require action and important results (abnormal or otherwise) should be highlighted. Values should also be easy to trend. Although some portals offer this feature, it is often difficult to find and use. Portals

should also be connected to trustworthy resources to facilitate users' learning about tests and their results, portals should be connected to.

5. Conclusion

Patients are no longer willing to settle for the maxim "no news is good news" when it comes to their lab results. There are many potential benefits when patients review their test results. Increased access and review of health data can translate to better patient engagement, improved health literacy, and in some cases, greater self-efficacy and self-management. Having access to results can also improve consumer satisfaction by reducing anxiety common when awaiting test results. It may also represent an important patient safety intervention by bringing attention to results that may have otherwise been overlooked. Knowing the motivation behind *why* people look at their online lab results and *what* they look for can influence the design of online lab results to better match the users' goals. For example, although we demonstrated that people want to see the trends of their values, this functionality is often limited or non-existent in online lab results.

References

[1] Directive 2011/24/EU of the European Parliament and of the Council of 9 March 2011 on the application of patients' rights in cross-border healthcare, 2011. http://data.europa.eu/eli/dir/2011/24/oj/eng (accessed January 19, 2022).

[2] Powers TL, and Bendall D. Improving Health Outcomes Through Patient Empowerment, *Journal of Hospital Marketing & Public Relations*. **15** (2004) 45–59. doi:10.1300/J375v15n01_05.

[3] Pillemer F, Price RA, Paone S, Martich GD, Albert S, Haidari L et al. Direct Release of Test Results to Patients Increases Patient Engagement and Utilization of Care, *PLoS ONE*. **11** (2016) e0154743. doi:10.1371/journal.pone.0154743.

[4] Shah SGS, Fitton R, Hannan A, Fisher B, Young T, Barnett J. Accessing personal medical records online: A means to what ends?, *International Journal of Medical Informatics*. **84** (2015) 111–118. doi:10.1016/j.ijmedinf.2014.10.005.

[5] Giardina TD, Baldwin J, Nystrom DT, Sittig DF, Singh H. Patient perceptions of receiving test results via online portals: a mixed-methods study, *Journal of the American Medical Informatics Association*. **25** (2018) 440–446. doi:10.1093/jamia/ocx140.

[6] Wilkie P. Patient views on understanding laboratory results, *Clinical Chemistry and Laboratory Medicine (CCLM)*. **57** (2019) 371–374. doi:10.1515/cclm-2018-0868.

[7] Video Conferencing, Web Conferencing, Webinars, Screen Sharing, *Zoom Video*. (n.d.). https://zoom.us/ (accessed May 7, 2021).

[8] Pernice K. Affinity Diagramming: Collaboratively Sort UX Findings & Design Ideas, *Nielsen Norman Group*. (2018). https://www.nngroup.com/articles/affinity-diagram/ (accessed March 29, 2021).

[9] Harrington HJ. Affinity Diagram, in: Techniques and Sample Outputs That Drive Business Excellence, Productivity Press, 2015.

[10] Monkman H, Macdonald L, Nøhr C, Tanaka JW, Lesselroth BJ. Hidden in Plain Sight: Overlooked Results and Other Errors in Evaluating Online Laboratory Results, *Studies in Health Technology and Informatics*. **290** (2021). doi:10.3233/SHTI290.

[11] Casalino LP, Dunham D, Chin MH, Bielang R, Kistner EO, Karrison TG, et al. Frequency of Failure to Inform Patients of Clinically Significant Outpatient Test Results, *Archives of Internal Medicine*. **169** (2009) 1123–1129. doi:10.1001/archinternmed.2009.130.

[12] Kelman A, Robinson C, Cochin E, Ahluwalia N, Braverman J, Chiauzzi E, et al. Communicating laboratory test results for rheumatoid factor: what do patients and physicians want?, *PPA*. **10** (2016) 2501–2517. doi:10.2147/PPA.S104396.

Challenges of Trustable AI and Added-Value on Health
B. Séroussi et al. (Eds.)

doi:10.3233/SHTI220540

The Rising Importance of e-Health in Norway

Rolf WYNN, [a,1] Vicente TRAVER SALCEDO,[b] Gunnar ELLINGSEN[a]
[a] *UiT The Arctic University of Norway, Tromsø, Norway*
[b] *SABIEN-ITACA, Universitat Politècnica de Valencia, Valencia, Spain*

Abstract. Drawing on three central sources of data on the development in e-health use in Norway (studies from the Norwegian Centre for e-Health Research, studies from Statistics Norway, and the Tromsø 7 Study), we describe the rising importance of e-health. Originally restricted to a limited use within the health services, in recent years the use of e-health has gained momentum both in the general population and within the traditional health services, as the Internet has offered easy access to health information as well as a range of other health-related services.

Keywords. E-health, statistics, population-based studies, Norway

1. Introduction

The importance of e-health to individuals and to health services is rising rapidly worldwide as well as in Norway [1-3]. The Covid pandemic has further strained health services in many countries already struggling with the demands of an increasingly elderly population. In addition to the obvious advantage of reducing person contact during an epidemic, e-health can increase access to services – which may be especially important concerning remote and otherwise underserved populations. Improving patient engagement by focusing on the preventative aspects of life-style changes is also an important feature of many e-health services, which also may alleviate the burden on strained traditional health services.

Even before most Norwegians had easy access to the Internet, telemedicine and e-health played a small but important role in the Norwegian health services [4]. The role of telemedicine and e-health was especially important in the more remote and sparsely populated parts of Norway, which had less access to specialized hospital-based services [4]. In addition to videoconferencing, the remote assessment of, for instance, electrocardiograms, retinal scans, x-rays, and pathology samples, have long been important for the health services in the more peripheral parts of the country [4].

As Norwegians have gained access to the Internet and smartphones have become commonplace, new services have been developed. While the most frequent e-health activity has been searching for health information online, other Internet-based services have recently become increasingly important [5].

[1] Rolf Wynn, Department of Clinical Medicine, Faculty of Health Sciences, UiT The Arctic University of Norway, N-9037 Tromsø, Norway; E-mail: rolf.wynn@uit.no.

In this paper, we will draw on three sources of data on e-health use development in Norway and discuss major developments in the use of e-health in Norway.

2. Methods

We briefly present and discuss data from three central sources on the development of e-health use in Norway; studies that have been performed by the Norwegian Centre for e-Health Research [6-8], studies conducted by Statistics Norway [9] and the population-based Tromsø 7 study [4,10].

The Norwegian Centre for e-Health Research carried out a series of surveys on e-health use in the Norwegian population during the first two decades of the millennium, including in 2000, 2001, 2003, 2005, 2007, and 2013 [6-8]. These surveys were performed on statistically representative samples of ca. 1000 respondents from the general population and carried out with the help of computer-assisted telephone interviews [6-8].

The Norwegian statistics agency, Statistics Norway, has since 2006 performed regular surveys to examine the use of e-health in the Norwegian population. These surveys have been based on statistically representative samples of 2000 respondents [9].

The third source of information regarding the use of e-health in Norway is the most recent (7^{th}) version of the population-based Tromsø Survey. This survey included people aged 40 or above living in Tromsø, a city of approximately 70,000 inhabitants, located in North Norway [10]. This study was conducted in 2015-2016, and involved 18,497 respondents, which was 65% of those invited. In addition to completing questionnaires (including on the use of e-health), the participants were subjected to a range of lab tests and physical examinations at a study-center [10].

3. Results

The Norwegian Centre for E-health Research has performed a number of surveys on the use of e-health in Norway. In some of the earlier studies from the beginning of the millennium, the use of the Internet for health purposes was low compared to present day figures. For instance, in 2000, 19% of the population used the Internet for health purposes, increasing to 31% in 2001 and 33% in 2002 [6]. However, there was a sharp increase in this period, and a more extensive European study including Norwegian respondents showed that the use of the Internet for health purposes rose to 59% in 2007 [7], and a few years later, in 2013, 78% of Norwegians had used the Internet for health purposes [8].

While the studies from Statistics Norway have yielded slightly lower figures than those from the Norwegian Centre for e-Health Research, the number of people that have used the internet to search for health-related information has increased similarly steeply, from 33% in 2006 to 76% in 2021 [9].

A recent version of the survey from Statistics Norway has also demonstrated that the use of online health services has expanded beyond information-seeking. For instance, in 2020, 41% used the Internet to gain access to their personal health information (read their medical record online), 33% made doctors' appointments online, and 28% used other health-related Internet-based services [9].

A main finding in the Tromsø 7 study was that as many as 52.7% of the population 40 years and older had used the Internet for health purposes during the last year [10]. Searching the Internet for health information was still the most frequent e-health activity – and 49.2% (10106/20252) had used search engines like Google for this purpose. However, other types of online services are gaining ground, and 13.5% (2687/19926) had used apps in the last year for information and advice on health and disease issues, 7.3% (1421/19481) had used social media like Facebook for these purposes, and 5% (969/19418) had used video services such as YouTube [4].

4. Discussion

The three data sources, the Norwegian Centre for e-Health Research-based studies [6-8], the studies from Statistics Norway [9], and the Tromsø 7 study [4,10], all show that e-health now is widely used in the general population in Norway. E-health use has developed over the last two decades, as the Internet gradually has become more accessible to more people, especially with the spread of smartphones. In parallel, a range of e-health services have been developed. More recently, apps, social media and video have taken their place in the realm of e-health.

The data show that the general use of the Internet for health purposes is now quite high in Norway. Compared to other European countries, Norway and the other Nordic countries have among the highest e-health usage [7,11,12]. However, the use of e-health in Norway is still mainly for obtaining information about health and illness. Googling symptoms as part of a decision-making process regarding whether to see a health professional is commonplace. If the process results in a doctor's visit, the information obtained from the Internet may be a central, although not necessarily outspoken, factor in the consultation [13]. Internet searches involving lifestyle advice, including exercise and dieting, are popular as is advice relating to complementary and alternative medicine. While the Internet is a very useful source of health-related information for most, some users experience challenges related to, for instance, inadequate searching abilities, trouble separating low-quality, erroneous or biased information from high quality information, misunderstanding information, too much information, or not being able to utilize the information [14].

Some of the e-health services widely implemented within the established public Norwegian health services are electronic health records (often also accessible by patients), videoconferencing for meetings and consultations (both within primary care and in hospital-based services), electronic prescriptions, and the online booking of consultations.

As technology develops, new types of e-health services are gaining ground. Many such services are marketed directly to consumers without any connection to the public or private health services (typically based on mobile apps or web-pages), and there is no overall system in place for quality assurance or for making use of the data in the health services. Other services have been developed by the health services themselves or in partnerships, and have become better integrated into the daily work in the health services.

Some treatments, especially within the behavioral field, are now available completely or in part as e-health interventions. While not yet widely available within the public health services, these and other innovations are likely to play an increasing part in health care in Norway in the future.

The three types of studies discussed here are all methodologically sound and have all been based on relatively large, statistically representative samples. One methodological difference between the studies is that those from Norwegian Centre for e-Health Research and Statistics Norway are based on phone-surveys, which was not the case with the Tromsø 7 study. While phone-based surveys can give a good indication of the use of e-health, these surveys may have a methodological weakness in that there are many who refuse to participate – which means that the sampling of the participants may be biased towards certain groups of people. Another difference is that whereas the studies from Norwegian Centre for e-Health Research and Statistics Norway had made a random stratified selection of people from the adult Norwegian population in general, the Tromsø 7 study invited all inhabitants of the municipality of Tromsø in the surveyed age range (40 years or more) to participate, and a large part [65%] did so.

The latest figures in the Norwegian Centre for e-Health Research-studies [8] (78%) and Statistics Norway [9] (76%) are therefore not directly comparable to the figure from the Tromsø 7 study [4] (52%), as the first two include all adult age groups and the latter only those 40 or above. We know that age is one of the main variables influencing the use of e-health [10], and middle aged are approximately four times as likely to use the Internet for health purposes as senior citizens [10]. One main advantage of the study design of large population-based surveys such as the Tromsø Studies is that a large amount of data is sampled from a high number of people. This gives high statistical power, allowing for the opportunity to examine the importance of variables that other designs might not.

Other important and statistically significant predictors of e-health use are gender (women use more), marital status (single use more), educational level (highly educated use more), income (high earners use more), chronic illness and disability (the ill and disabled use more), and consulting the GP in the last year (use more) [10]. Conversely, some groups have a low use of e-health, possibly because of a lower interest in the topic in general. However, some groups – such as the oldest or most ill – may not be able to use the Internet or find it too difficult to use. Others may be unaware of these services or may simply not trust them. There still exists a 'digital divide' in e-health in Norway. However, this concept does no longer relate to physical access to the Internet. Today, almost all Norwegians have physical access through their smartphones, pads and computers. Instead, the present day digital divide relates primarily to a choice made not to use the available technologies.

5. Conclusions

The use of e-health in Norway has been increasing during the last two decades. Today, most of the population – including a large part of the elderly population- use e-health services. While using search engines to find information about health and illness remains the most frequent e-health activity, apps, social media, and video are becoming more important sources. Moreover, technological innovations as well as increased use of e-health within the public health services are paving the way for the even more widespread use of e-health in the future.

The Covid-19 pandemic has increased the interest in and use of e-health technologies in Norway and around the world. The importance of e-health to people in general and to the health services is likely to increase further as users are becoming more accustomed to and discovering the benefits of e-health.

E-health services that draw on important qualities such as ease of access, patient-centeredness, and the best outcomes from the patients' perspective, may, in line with the concepts around value-based health care, be strongly suited for our future health care [15].

References

[1] Wynn R, Gabarron E, Johnsen JAK, Traver V. Special issue on e-health services. Int J Environ Res Public Health. 2020;17:2885.
[2] Wynn R. E-Health in Norway before and during the initial phase of the Covid-19 pandemic. Stud Health Technol Inform. 2020;272:9-12.
[3] Acharibasam JW, Wynn R. Telemental health in low- and middle-income countries: A systematic review. Int J Telemed Appl. 2018;2018:9602821.
[4] Wynn R, Ellingsen G. The master study in telemedicine and e-health at the University of Tromsø, Norway, 2005-2018. Stud Health Technol Inform. 2020 Nov 23;275:217-221.
[5] Yigzaw KY, Wynn R, Marco-Ruiz L, et al. The association between health information seeking on the Internet and physician visits (The Seventh Tromsø Study - Part 4): Population-based questionnaire study. J Med Internet Res. 2020;22(3):e13120.
[6] Wangberg S, Andreassen H, Kummervold P, et al. Use of the internet for health purposes: trends in Norway 2000-2010. Scand J Caring Sci. 2009;23(4):691-696.
[7] Kummervold PE, Chronaki CE, Lausen B, Prokosch HU, Rasmussen J, Santana S, et al. eHealth trends in Europe 2005-2007: a population-based survey. J Med Internet Res. 2008;10(4):e42.
[8] Sørensen T, Andreassen HK, Wangberg SC. E-helse i Norge 2013. Nasjonalt senter for samhandling og telemedisin,Universitetssykehuset Nord-Norge, 2013.
[9] Statistics Norway. Use of ICT in households. Available from: https://www.ssb.no/statbank/table/06998/tableViewLayout1/
[10] Wynn R, Oyeyemi SO, Budrionis A, Marco-Ruiz L, Yigzaw KY, Bellika JG. Electronic health use in a representative sample of 18,497 respondents in Norway (The Seventh Tromsø Study - Part 1): Population-Based Questionnaire Study. JMIR Med Inform. 2020;8(3):e13106.
[11] Kummervold PE, Wynn R. Health information accessed on the internet: the development in 5 European countries. Int J Telemed Appl. 2012;2012:297416.
[12] Living Report of the WHO Symposium on the Future of Digital Health Systems in the European Region, Copenhagen, Denmark, 6–8 February 2019. Copenhagen: WHO Regional Office for Europe; 2019.
[13] Van Riel N, Auwerx K, Debbaut P, Van Hees S, Schoenmakers B. The effect of Dr Google on doctor-patient encounters in primary care: a quantitative, observational, cross-sectional study. BJGP Open. 2017;1(2):bjgpopen17X100833.
[14] Wynn R. Challenges related to health information on the Internet: a narrative review. Int J Integr Care. 19(S1): A391.
[15] Seoane F, Traver V, Hazelzet J. Value-driven digital transformation in health and medical care. In: Fernandez-Llatas C. (eds) Interactive process mining in healthcare. Health Informatics. Springer, Cham, 2021.

Challenges of Trustable AI and Added-Value on Health
B. Séroussi et al. (Eds.)

doi:10.3233/SHTI220541

Disparities in Regional Publication Trends on the Topic of Artificial Intelligence in Biomedical Science Over the Last Five Years: A Bibliometric Analysis

Kirubel Biruk SHIFERAW[a,1], Dagmar WALTEMATH[a] and Atinkut ZELEKE[a]
[a] *Department of Medical Informatics, Institute for Community Medicine, University Medicine Greifswald, Germany*

Abstract. Bibliometric analysis is a scientific method that allows researchers to explore the current trend in a certain research area using citation information. This study aims to provide a meta-view of artificial intelligence studies focused on biomedicine in the last five years, which will provide an insight into current trends and future research directions. Besides the observation of increased publication rates in the area of AI in biomedicine, the results indicate a lower contribution from and a sparser network connectivity of countries with limited resources. Thus, working toward collaboration in terms of infrastructure and implementing alternative solutions such as FAIR (Findable, Accessible, Interoperable and Reproducible) and open access platforms could improve the collaborative nature of international health projects.

Keywords. Artificial Intelligence, Medicine, Biomedicine, Bibliometric Analysis

1. Introduction

Leveraging the vastly increasing volume of medical data being generated, the application of advanced Big Data analysis approaches such as Machine learning (ML) and Deep Learning (DL) are considered important milestones of the upcoming requirements in healthcare [1]. Several studies flagged the relevance of applying artificial intelligence (AI) in risk, treatment and event/outcome prediction [2]. The number of studies published in peer reviewed academic journals and suggesting customized models for predictions based on Big Data is steadily increasing. Synthesizing evidence in this area of research will provide a comprehensive input for upcoming clinical and methodological guideline development and an indication for potential research areas.

Bibliometric analysis is a scientific method that enables researchers to explore the current trend in a certain research area using citation information [3]. It mainly provides a bird's-eye view on the activities in the research domain, showing who is doing what, where, with whom and the intensity of cross-country, author and affiliation collaboration

[1] Corresponding Author, Kirubel Biruk Shiferaw, Walther-Rathenau-Str. 48, Department of Medical Informatics, University Medicine Greifswald, D-17475 Germany; E-mail: s-kishif@uni-greifswald.de.

on that specific research area. This study aims to provide a meta-view of artificial intelligence studies focused on biomedicine in the last five years, which will provide an insight for further research and collaboration.

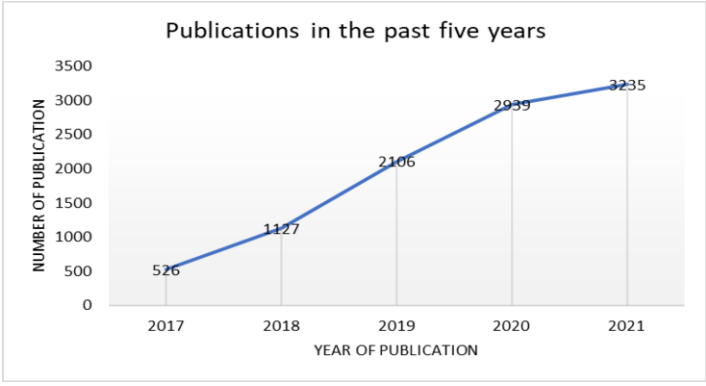
2. Methods

After keywords had been identified from previous literature ("Artificial intelligence" OR "Machine intelligence" OR "artificial neutral network*" OR "Machine learning" OR "Deep learn*" OR "Natural language process*" OR "Robotic*" as a major intervention) [4], an iterative search was conducted on PubMed, which can be used as a bibliographic source for biomedical search containing more than 33milion citations [5]. Medical subheadings, mixed keywords, truncated keywords and controlled vocabularies were used. The search was limited to human species, English language, studies with abstract and publications since 2017 (inclusion criteria). Using bibliometric analysis (bibliometrix package) [3], we then explored trend, growth rate and pattern of publications as well as collaboration networks.

3. Results

The search resulted in a total number of 24,979 studies. The first 10,000 best match studies were considered for further bibliometric analysis. The studies were identified using the PubMed built-in Best Match algorithm which analyzes each citation based on how many search terms are found and in which fields they are found [6].

Figure 1: Annual publication trend on the area of AI in Biomedical science 2017-2021



The result indicated that the majority of studies were journal articles (40.9% / 4,087). The annual percentage growth rate was 57.48% between 2017 and 2021. The highest increase in publication was observed between 2018 and 2020; it accounted for 77.45 % of the increase in scientific contributions in AI focusing on biomedicine in the last five years. The result includes a list of the top five relevant sources, corresponding author countries and affiliations (Table 1). The journal *Scientific Reports* was identified as the most relevant source for publications. The USA and the University of California were

also identified as the most productive country and the most relevant affiliation, respectively. Regarding collaboration, Germany was identified as the most collaborative country with 40.4% of publications, followed by the USA (29.6%).

Table 1: Top five sources, corresponding authors' countries and affiliations associated with scientific contributions focused on AI in biomedical science since 2017.

Top five sources	Top five corresponding authors' countries, SCP and MCP		Top five relevant affiliations
	Country (Number of publications)	SCP (%), MCP (%)	
Scientific reports (469)	USA (1765)	1243 (70.4%), 522 (29.6%)	University of California (1091)
Sensors (Basel Switzerland) (433)	China (1576)	1157 (73.4%), 419 (26.6%)	Stanford University (837)
Plos one (347)	Korea (512)	443 (86.5%), 69 (13.5%)	Harvard Medical school (676)
Annual international conference of the IEEE Engineering in medicine and biology society (306)	Germany (319)	190 (59.6%), 129 (40.4%)	SUN YAT-SEN University (396)
IEEE journal of biomedical and health informatics (197)	Japan (310)	272 (87.7%), 38 (12.3%)	Renim Hospital of Wuhan University (347)

SCP: Single country publication, MCP: Multiple country publication

Figure 2 shows each country's scientific contribution in the last five years. Considering the top contributor in each continent, USA (n=1,754), China (n=1,575), Germany (n=319), Australia (n=195), Brazil (n=69) and Egypt (n=23) are found to be the top contributors in their respective continent.

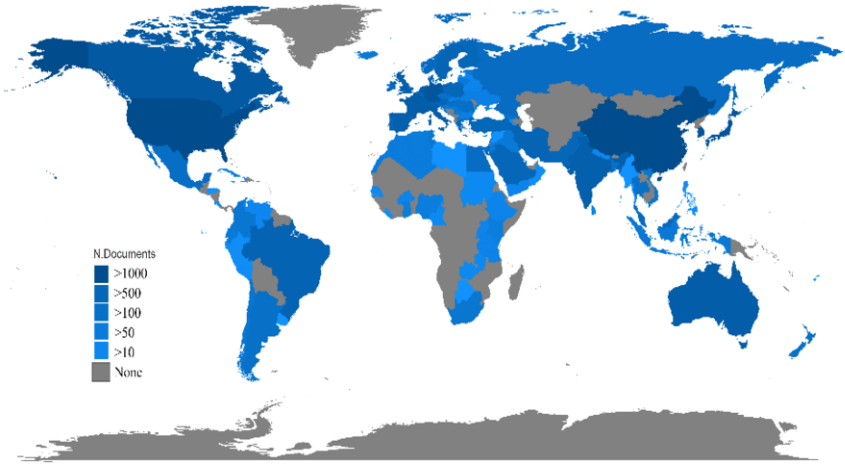


Figure 2: Country's scientific contribution

The country collaboration network depicted in Fig. 3 indicates a very intense and large number of collaborations among higher income countries but less intense networks between low and high-income countries. Moreover, a very low or no collaboration network is observed among low-income countries.

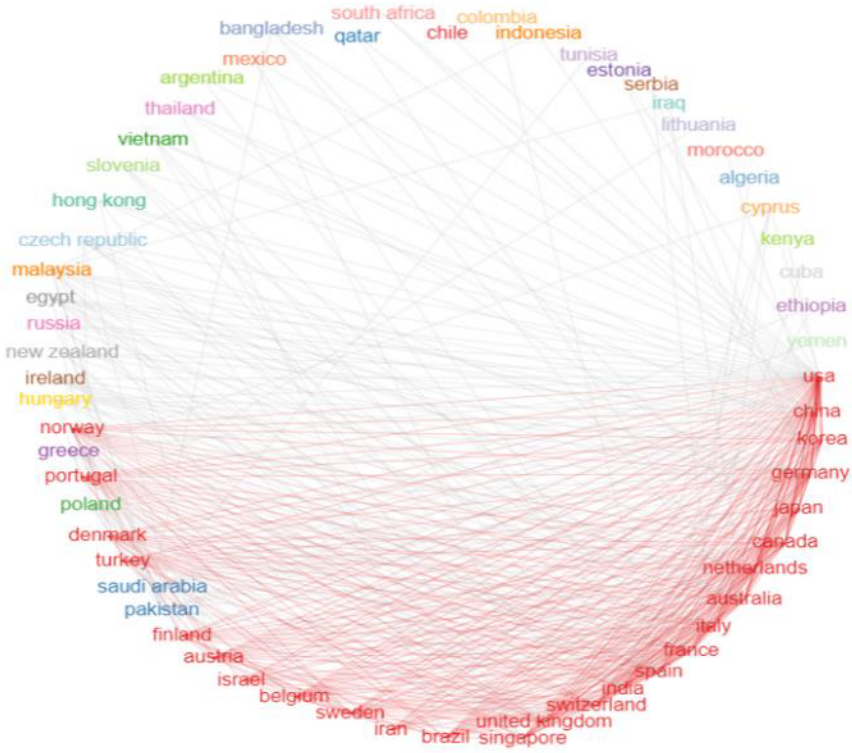


Figure 3: Country's collaboration network

4. Discussion

As a result of the ever-growing computing and storage capacity, the number of studies applying AI in biomedicine has drastically increased in the past decade [7]. More specifically, we noticed a significant increase in the number of publications in the past five years and especially after 2019. This could be due to the rush for digital solutions in early diagnosis, treatment and drug/vaccine development for the COVID-19 pandemics that we are still dealing with since 2019 [8, 9].

In terms of scientific contribution in the last five years, nearly half (44.8%) of the scientific contributions are from five countries only (USA, China, Korea, Germany and Japan), and the contribution from countries in Africa and central Asia is low in number and also with respect to collaboration networks. This could be due to the significant gap in national research and development infrastructure among the high- and low-income countries. AI-related research is dependent on data and computing power. Lack of infrastructural capacities for data, Internet, Hardware or Software in Central Asia and Africa could be one reason for non-uniform international collaborations [10]. This finding is also in line with another study by *Tran, B.X., et al, 2019* which indicated an intense collaboration among Europe [7].

To achieve the 2030 United Nations Sustainable Development Goals, specifically SDG-9 (Target 9.5) and SDG-17 (Target 17.6 and 17.9), improving North-South, South-

South and triangular regional and international cooperation on knowledge sharing, access to science, technology and innovation is important [11].

The current study comprises only published results in English language and this could result in publication bias since there are number of publications and grey literatures available in local databases in different languages across the world.

5. Conclusions

Besides the mounting indication of increased publication in the area of AI in biomedicine in the last five years, our results clearly show a lower contribution and networking activity from resource limited countries. Thus, further collaboration in terms of infrastructure development and implementation of solutions for data sharing and interoperability should be a mid-term goal in the field of AI. Here, FAIR concepts and open data platforms can help to improve the necessary networking capabilities and enable more intense collaborative data exchange.

References

[1] Matsangidou M., et al. Machine Learning in Pain Medicine: An Up-To-Date Systematic Review. Pain and therapy, 2021. 10(2): p. 1067-1084.
[2] Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. Nature medicine, 2019. 25(1): p. 44-56.
[3] Aria M, Cuccurullo C. bibliometrix: An R-tool for comprehensive science mapping analysis. Journal of informetrics, 2017. 11(4): p. 959-975.
[4] Tran BX, et al. Global Evolution of Research in Artificial Intelligence in Health and Medicine: A Bibliometric Study. Journal of Clinical Medicine, 2019. 8(3): p. 360.
[5] Falagas ME, et al. Comparison of PubMed, Scopus, web of science, and Google scholar: strengths and weaknesses. The FASEB journal, 2008. 22(2): p. 338-342.
[6] Fiorini N, et al. Best Match: New relevance search for PubMed. PLoS biology, 2018. 16(8): p. e2005343-e2005343.
[7] Tran BX, et al. Global Evolution of Research in Artificial Intelligence in Health and Medicine: A Bibliometric Study. 2019. 8(3).
[8] Monteleone S, et al. Fighting COVID-19 with Artificial Intelligence, in Artificial Intelligence in Drug Design. 2022, Springer. p. 103-112.
[9] Al-Hashimi M, and Hamdan A. The Applications of Artificial Intelligence to Control COVID-19, in Advances in Data Science and Intelligent Data Communication Technologies for COVID-19. 2022, Springer. p. 55-75.
[10] Cary M, Rockwell T. International Collaboration in Open Access Publications: How Income Shapes International Collaboration. Publications, 2020. 8(1): p. 13.
[11] Nations U. 2030 Sustainable Development Goals (SDGs) https://sdgs.un.org/goals. [cited 2022 01/24/2022.

Challenges of Trustable AI and Added-Value on Health
B. Séroussi et al. (Eds.)

doi:10.3233/SHTI220542

Design for a Virtual Peer-to-Peer Knowledge to Action Platform for Type 2 Diabetes

Elham GOLKHANDAN[a], Alessia PAGLIALONGA[b], Aziz GUERGACHI[a], Marie-Therese LUSSIER[d], Claude RICHARD[d], Laurette DUBE[e], Ian ZENLEA[f], Robert KYBA[g], Manmohan MITTAL[h], Warren "Smokey" THOMAS[i], Karim KESHAVJEE[h,1]

[a] *Ryerson University, Toronto, Canada,* [b] *Institute of Electronics, Information Engineering and Telecommunications (IEIIT), National Research Council of Italy (CNR), Milan, Italy,* [c] *Department of Mathematics and Statistics, York University, Toronto, Canada*, [d] *Universite de Montreal, Montreal, Canada,* [e] *McGill Centre for the Convergence of Health and Economics, McGill University, Montreal, Canada,* [f] *Trillium Health Partners, Mississauga, Canada,* [g] *Strategic Global Counsel, Toronto, Canada,* [h] *Institute of Health Policy, Management and Evaluation, University of Toronto, Canada,* [i] *Patient Representative, Kingston, Ontario*

Abstract. Many patients with Type 2 Diabetes (T2D) have difficulty in controlling their disease despite wide-spread availability of high-quality guidelines, T2D education programs and primary care follow-up programs. Current diabetes education and treatment programs translate knowledge from bench to bedside well, but underperform on the 'last-mile' of converting that knowledge into action (KTA). Two innovations to the last-mile problem in management of patients with T2D are introduced. 1) Design of a platform for peer-to-peer groups where patients can solve KTA problems together in a structured and psychologically safe environment using all the elements of the Action Cycle phase of the KTA framework. The platform uses Self-Determination Theory as the behavior change theory. 2) A novel patient segmentation method to enable the formation of groups of patients who have similar behavioral characteristics and therefore who are more likely to find common cause in the fight against diabetes.

Keywords: Knowledge To Action (KTA) Framework, Diabetes Education, Behavioral change theory, patient segmentation, peer-to-peer education.

1. Introduction

The Knowledge to Action (KTA) Framework (Figure 1) is an evidence-informed ontology for making knowledge applicable to the daily life of patients with a variety of health conditions. This framework consists of two components: The Knowledge Creation Pyramid and the Action Cycle, each of which comprises multiple elements (Figure 1). Increasingly it is apparent that the action phase or 'last-mile' of adapting knowledge to patients' needs is sorely lacking in many patient education programs [1, 2]. Conceptual frameworks, such as the KTA, are recommended as tools to support the application of theories and models in the Action Cycle [1]. Recent conceptual analytic work in our lab shows that patient empowerment cannot occur through empowerment of individuals

[1] Corresponding Author, E-mail: karim@infoclin.ca

Figure 1. Knowledge to Action (KTA) Framework (Graham et al., 2006, with permission)

alone. Empowerment needs to occur at 3 levels –individual [3], communities and groups and the societal level [4].

The primary objective was to design a scalable reference architecture for a digital platform that incorporates all aspects of the KTA Action Cycle, can incorporate behavior change theory, principles of collective intelligence, behavioral economics, proven behavior change interventions and healthcare marketing to help patients with Type 2 diabetes (T2D) improve control of their disease by developing skills for healthy behaviors and habits. To our knowledge, this is the first description of a digital platform that incorporates all elements of the Action Cycle of the KTA framework. It is also the first digital KTA framework to incorporate a peer-to-peer component.

2. Methods

We initially developed a draft architecture for a peer-to-peer platform (P2PP) that can engage patients to take translated knowledge and convert it into action. The draft P2PP starts by extracting data from physician electronic medical record (EMR) systems, identifying patients with poor diabetes control, segmenting them into medication-taking behavioral segments using a k-means clustering algorithm [5], engaging them and encouraging them to enroll into a healthy skills development workshop and reporting the outcomes back to the physician.

Subsequently, we mapped elements of the Action Cycle of the KTA framework onto the draft P2PP and added new functions to accommodate elements of the cycle not included in the initial draft version. We iterated the P2PP until we obtained consensus on the feasibility and efficiency of the workflow that contained all elements of the KTA Action Cycle. We also mapped the Self-Determination Theory of behavior change [6, 7] onto the architecture and identified where behavioral segmentation could be added. Privacy and security are handled elsewhere and was not in scope for this project.

3. Architecture for KTA

The final architecture (Figure 2) is a closed loop to ensure that patients receive appropriate follow-up from their primary care providers and that continuity of care is preserved. Patients with poor control of their diabetes (e.g., uncontrolled blood pressure, cholesterol and/or hemoglobin A1c), are segmented (using a k-means clustering) to identify different behavioral segments (lifestyle, cholesterol, basic medication and extensive medication treatment preference groups), described elsewhere [5].

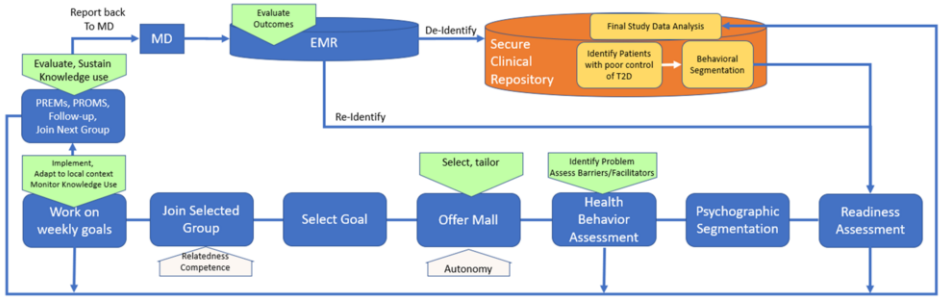


Figure 2. Reference Architecture for Peer-to-Peer KTA Diabetes Workshops

Then, patients are administered validated questionnaires to assess: readiness for change, health behaviors, health literacy and a psychographic (attitudinal) segmentation questionnaire. The psychographic segmentation and health literacy assessment are used to tailor the language of subsequent communications to the preferences of the patient. The health behaviors assessment is where the Action Cycle elements begin (lime green) [8]. At this stage, problems (why is the patient's disease out of control?) and barriers and facilitators to correct the problems are identified. Subsequently, the patient is offered a choice of lifestyle programs (diet, exercise, stress management, symptom management or medication adherence program) that is tailored to their individual health behaviors and disease control (The Mall).

The Mall allows the patient to select the intervention of their choice, promoting autonomy in line with Self-Determination Theory (SDT, cream) and supports selection and tailoring, as described in the Action Cycle. Once the patient has set their goal, they are enrolled into a workshop where they meet other patients with similar characteristics. The workshop is a 12-week peer-to-peer skills development, problem solving and relationship building program, consistent with the Relatedness and Competence elements of SDT (cream arrow). The workshop participants meet weekly, engage each other to implement micro-goals, measure and display their collective progress and help each other find ways of adapting materials to their local context, consistent with the elements of the Action Cycle.

Once the patients have completed the workshop, they are asked to complete regular post-workshop questionnaires and patient reported measures. A report is generated and sent to the physician so that they are kept informed and can reinforce participation and encourage other patients to adopt the program. Finally, data from the patient's visit with their physician is used to assess the impact of the program on diabetes control.

4. Benefits of the P2PP architecture

This P2PP reference architecture has several characteristics which distinguish it from other similar projects. (i) Integration into existing EMRs helps to identify all patients that have poor control of their disease. Population-level data makes it easier to identify missing data, increases transparency and decreases risk of bias, unlike when patients are approached individually and opportunistically in a clinic. Integration with EMRs enables scalability and integration into the healthcare system. (ii) The Action Cycle of the KTA Framework adds an evidence-informed design component to the P2PP that is rarely included in similar projects. Addition of the Self-Determination Theory explicitly includes functions that support a behavior theory. (iii) Unlike current one-size-fits-all education programs, the P2PP is designed to suit each T2D patient based on their individual preferences, attitudes and habits; e.g. language, diet, exercise, sleep, stress, and regardless of sex, gender, social or ethnic characteristics. Behavioral and psychographic patient segmentation enable tailoring of communications and personalization of options for workshops. (iv) Patients classified in the same segment who share similar preferences, outlooks and constraints but have distinct problem solving and self-motivation capabilities work together to achieve collective goals that might be difficult to achieve individually. Utilizing positive deviance and experience sharing among individuals of the same group, the P2PP enables patients to learn from peers who have better control of their disease and helps them develop healthier attitudes and behavior change skills in an environment that supports relationship building and group motivation. (v) The P2PP enables person-centered decision-making by creating a choice architecture that maximizes opportunities for improvement of disease control. This approach increases scalability, sustainability and patient engagement by enhancing autonomy in the individual's natural environment and augments existing educational programs and physician-mediated recommendations. (vi) The reference P2PP architecture is scalable to additional interventional components such as additional behavior change theories, mobile apps, explainable artificial intelligence, nudges and other behavior change techniques. (vii) The P2PP architecture enables patients to join additional workshops to develop additional skills whenever they are ready to do so. This addresses the issue of one-off interventions which make an impact, only for patient behavior to revert to previous behaviors after the intervention is over. Long-term behavior change needs consistent and on-going intervention. (viii) Finally, the reference P2PP architecture proposed is technology agnostic and can be developed using cloud-based no-code and low-code applications readily available on the web which helps address the problem of affordability and sustainability.

5. Conclusion

To our knowledge, this peer-to-peer platform for translating knowledge into action is the first of its kind that incorporates an evidence-informed ontology and provides T2D patients a platform that is designed based on their own behavioral and psychographic segments, preferences and socioeconomic contexts. This approach encourages patients to change their behavior towards a healthy lifestyle by looking at and following peers with better control of biomarkers as role models and thus challenges current group interventions which do not use segmentation, and which have clinicians or experts as facilitators.

The platform can also be adapted for other diseases (e.g. cardiovascular diseases, dementia and cancer) that can be improved through lifestyle change. By leveraging individual attitudes and beliefs in a group-centered, patient-led way, this new system is designed to contribute to empowering patients at the group and community level and transforming the healthcare system from a reactive to proactive stance.

References

[1] Field B, Booth A, Ilott I, Gerrish K. Using the Knowledge to Action Framework in practice: a citation analysis and systematic review. Implementation Science 2014, 9:172.

[2] Graham I, Logan J, Harrison M, Straus S, Tetroe J, Caswell W, et al. Lost in knowledge translation: time for a map? J Contin Educ Health Prof. 2006, 26: 13-24. 10.1002/chp.47

[3] Lussier MT, Richard C, Binta Diallo F, Boivin N, Hudon C, et al. I am ready to see you now, Doctor! A mixed-method study of the Let's Discuss Health website implementation in Primary Care. Health Expect. 2021 Apr;24(2):243-256. doi: 10.1111/hex.13158. Epub 2020 Dec 7. PMID: 33285012; PMCID: PMC8077096.

[4] Daruwalla Z, Thakkar V, Aggarwal M, Kiasatdolatabadi A, Guergachi A, Keshavjee K. Patient Empowerment: The Role of Technology. Stud Health Technol Inform. 2019;257:70-74. PMID: 30741175.

[5] Paglialonga, A., Theal R, Barber D, Kyba R, Guergachi A, Keshavjee K. Patient Segmentation in Primary Care for Educational Interventions in Type 2 Diabetes. Manuscript in preparation.

[6] Phillips AS, Guarnaccia CA. Self-determination theory and motivational interviewing interventions for type 2 diabetes prevention and treatment: A systematic review. J Health Psychol. 2020 Jan;25(1):44-66. doi: 10.1177/1359105317737606. Epub 2017 Nov 9. PMID: 29119829.

[7] Teixeira PJ, Carraça EV, Markland D, Silva MN, Ryan RM. Exercise, physical activity, and self-determination theory: a systematic review. Int J Behav Nutr Phys Act. 2012 Jun 22;9:78. doi: 10.1186/1479-5868-9-78. PMID: 22726453; PMCID: PMC3441783.

[8] Lussier MT, Richard C, Glaser E, Roberge D. The impact of a primary care e-communication intervention on the participation of chronic disease patients who had not reached guideline suggested treatment goals. Patient Educ Couns. 2016 Apr;99(4):530-541. doi: 10.1016/j.pec.2015.11.007. Epub 2015 Nov 12. PMID: 26657041.

Challenges of Trustable AI and Added-Value on Health
B. Séroussi et al. (Eds.)

doi:10.3233/SHTI220543

Supporting Digital Inclusion and Web Accessibility for People with Cognitive Disabilities

Ioannis POULTOURTZIDIS [a,1], Aikaterini Marina KATSOULI [a], Savvas ANASTASIADES [a], Stavroula MAKROGLOU [a], Efstathios SIDIROPOULOS [a] and Panagiotis D. BAMIDIS [a]
[a] *Lab of Medical Physics and Digital Innovation, School of Medicine, Aristotle University of Thessaloniki, Greece*

Abstract. The topic of digital inclusion and web accessibility for People with Cognitive Disabilities has increased over the COVID-19 pandemic times. The LIVE IT project is attempting to shed some light into this. This piece of work uses insights gained from focus groups and interviews that were held to assess the needs analysis and the existing knowledge gap of this societal problem. To this end, preliminary results of user engagement with digital tools and web services as well as their evaluation are presented herein.

Keywords. Inclusion, Web accessibility, Cognitive Disabilities, Living Labs, Co-creation

1. Introduction

The COVID-19 pandemic has brought into light the existing scarcity of people with cognitive disabilities (PwCD) in web accessibility [1][2] while the urge for digitalization in many aspects of daily life increased [3][4][5]. The definition of "web accessibility" indicates that websites, tools, and technologies should be developed and designed in a way that people with any aspect of cognitive impairment or deficit can use them [6]. Research shows that cognitive impairments bring limitations and restrictions in the conceptualization of information and usual activities [7], which increases reliance and dependency on other people, for instance, carers, to act on behalf of PwCD during quarantine. Although there has been some effort in digital design of technologies that can be beneficial to PwCD [8][9][10][11], differences in digital inclusion between sub-groups of diagnosis and impairments, for instance, autism, aphasia, ADHD (Attention deficit hyperactivity disorder), and relevant limitations in people with physical impairments, like visually impaired people, remain unresolved [12][13]. Co-creation sessions which include a wide range of participatory practices for design and decision making with stakeholders and users [14], may be a key tool that reflects the perspectives and voices of PwCD and provide new opportunities to tailor new services to the needs of the cognitive disability community. In this paper, we outline the methodology

[1] Corresponding Author, Ioannis Poultourtzidis, Lab of Medical Physics and Digital Innovation, School of Medicine, Aristotle University of Thessaloniki, Greece; E-mail: giannis.poultourtzidis@gmail.com.

followed in the LIVE IT project and we provide the first evidence from preliminary data collections that seem to shed some light into the web access problem of PwCD.

2. Concept building blocks

Living Labs, as a new model of co-creation design practice, is an innovative approach relying on intensive user involvement through co-creation, using real-life settings and a multi-stakeholder approach [15]. Co-labs are considered as interdisciplinary services centers aiming to enhance the provision of accessible online services without barriers and exclusions acting as areas of social innovation for the exchange of skills, resources, and results. The use of co-labs and Living Labs methodology holds a central role in the LIVE IT project and leads the co-design processes as a way of innovative and user-centered nature participatory approach. It is believed that such practices can upgrade the web accessibility of PwCD and offer insights on how existing tools may become beneficial for persons with a wider range of impairments.

3. Methods

Participant inputs were gathered through multi-center focus groups and interviews. An online assistive toolkit was developed to provide suggested web accessible tools and services to the participants while requesting their prompt evaluations during the co-creation sessions. The whole process was ethically approved by the Ethics Committee of the Aristotle University of Thessaloniki (Greece), no. of vote 247165/2021, chairperson: Dimitrios Stamovlasis.

3.1. Open Toolkit

The LIVE IT toolkit consists of three main tabs. Each tab focuses on a specific aspect of the project goals. It is an assistive tool not only for PwCD but for caregivers, helpers, family members as well as stakeholders. The first tab is the "Catalogue and Stakeholders", whose main purpose is to help the users of the toolkit locate the nearby stakeholders that may be of interest. This is highly important because PwCD and their caregivers or family members often try unsuccessfully to locate a stakeholder. The second tab is the "Advisor tool and Guidelines". On this tab, a user can find a complete list of available services and tools for specific tasks to complete which are sorted by the IT platform and the lists are filled with the output of the co-design scenarios that took place at partners' Labs. These lists will also be fed by Hackathons' output, where the scenarios were tested, and tools were evaluated by the users themselves. Another important feature of the advisor tool is that it provides ratings for each tool. The last tab is the "Online Community and Makerspace". It links all the social networks and Living Labs' communities to build up the online community of the project. Thus, every user feels connected to a wide network of peers and relevant stakeholders where assisting technologies can be helpful. The toolkit is available through a web browser, while it is highly portable and accessible from a vast variety of devices from smartphones to desktop PCs (Personal Computers). The toolkit is being constantly evaluated during the series of co-Labs' sessions, Hackathons as well as Makerspace activities.

3.2. Co-Creation Sessions

At the first stage, focus groups with observation processes or semi-structured interviews were held. Eleven students (adolescents and young adults between the age of 15-24) with cognitive difficulties as well as their teachers and various healthcare specialists participated, pointing out the difficulties they face in their interaction with digital technologies. As a next step, co-creation sessions were implemented, where 10 students with major cognitive disabilities (adolescents and young adults between the age of 15-24) with the help of their teachers interacted with the Open Toolkit. This is an ongoing process where data will be collected during the planned hackathons.

4. Results

The preliminary qualitative results of the co-creation sessions are presented in Table 1.

Table 1. The qualitative results by each co-creation session

Means of Data Collection	Indicative Results
Focus Group Session with people with cognitive disabilities and their caregivers (observation, focus group and semi-interviews)	(a) People with major cognitive disabilities (people in autism spectrum with comorbidities, cognitive disorders and down syndrome) struggle during their interactions with digital devices (b) They need constant help and guidance of caregivers and teachers during this interaction (c) The digital interaction framework needs to be as structured as possible, because they have difficulty in understanding abstract concepts such as the internet environment, etc. (d) Serious games can be beneficial. Participants were more familiar with this digital environment
Interview with caregivers of persons with dementia	(a) The persons with dementia could be benefited from danger detecting devices in their home environment and devices that alert caregivers or personal assistants or include panic buttons (b) The persons are not able to learn to interact with digital devices and media, because of memory loss. The use of internet and digital devices should be supported by their caregivers (c) Platforms like YouTube are more acceptable by the persons
Interview with caregivers of persons with cognitive disability	(a) There is intense "technophobia" and unfamiliarity in their family environment (b) PwCD could benefit from a digital assistant (c) PwCD like interacting with embossed surfaces and buttons (d) The persons with cognitive disabilities face great difficulty in figuring out how to operate devices
Interviews with people who have been diagnosed with ADHD and a caregiver of a person with ADHD	(a) Some devices for spelling, text-to-speech or speech-to-text services, digital calendars, note-taking and digital applications are used by people with ADHD in their digital interactions to accomplish their tasks and help them (b) Unrestricted internet use could worsen the symptoms of ADHD or may worsen the existed attention deficit of the people that have been diagnosed with ADHD (c) Online tools, however, have the benefit of aiding in completing the activities of individuals and making those activities more efficient

During the first of the hackathon series, the participants evaluated text and voice converters as well as voice recognition services of platforms to be more useful in the context of the project. Tools with autocorrection functions in words were not evaluated

as very functional. In most of the tools, the guidance and support of the teacher is crucial for the participants' interaction to be achieved. The results of the first hackathon are presented in Table 2.

Table 2. The results of the evaluation of the digital tools during the 1st of the hackathon series

Tools evaluated	1st Hackathon results
Azure Microsoft Speech to text	It has been proven useful, as it was sensitive to the detection of sounds and words and even when people did not pronounce the words clearly, but also because they provide the choice of their native language
Text-from-to-speech	This tool offers a lot of options (shrink, zoom, clear the content, etc.) which could be proven useful for people with comorbidities, such as visual impairment. Most of the students managed to make proper use of the voice command that the tool offers
Google Chrome voice function	The advantage in the use of that tool, as it is described by most of the students and their teachers is that it automatically displays the suggested web pages according to the spoken word
On-Screen-Keyboard (OSK)	This tool has not proven functional in the co-design / co-creation sessions because the vast majority preferred the conventional keyboard
Autocorrect in Windows 10	The tool made small corrections to written words. The students who participated reacted positively with it in co-working / co-design labs
Coloradd.net	Though amusing for students to interact with it and change colors and backgrounds of words, according to teachers it has not proven useful
Myaccessangel.com	It was used under the guidance of teachers. Most of the students who took part in this session used its various options, such as changing the font or font size, but according to the teachers did not seem to like it
Contrastchecker.com	Students who took part in the sessions used that tool under the guidance of the teachers and they manage to make color changes both in the "background" and in the "foreground" of the words. According to their teachers, however, this tool has been proven a useful asset for the students in the case of completing suggested scenarios

5. Conclusions

This paper presents the first insights of the LIVE IT project's results which are expected to expand in the next few months. The role of teachers or caregivers in the interaction of people with cognitive and neurodevelopmental disabilities plays a central role in their life. The difficulties that they and their caregivers face when using digital technologies were highlighted during the interviews analyzed herein. The simple interface of the Toolkit and the structured working scenarios where participants could easily interact was proven to be a useful methodological (good) practice. Due to the problems that people with cognitive disabilities face in the production of written and spoken language, digital tools and applications could probably be used to assist them in their life. These findings are related to the sample of participants in the collaboration sessions, who were people with neurodevelopmental disorders, such as cognitive disability, people on the autism spectrum with comorbidity, and people with Down syndrome, while the data were collected in the first half of the project. The results will be complemented with new insights and knowledge generated by the series of hackathons and webinars planned by the partnership over the next few months of the project realm.

Acknowledgements

This research has been partially funded by the European Commission under the LIVE IT project (LC-01624438/101018048). Authors would also like to thank the rest of the LIVE IT consortium, namely Arcola Research LLP, National University of Ireland Galway and Universidade Católica of Portugal.

References

[1] Blanck P. (2014). eQuality: Web accessibility by people with cognitive disabilities. Inclusion, 3(2), 75-91. https://doi.org/10.1352/2326-6988-3.2.75

[2] Chadwick D, Wesson C, Fullwood C. (2013). Internet access by people with intellectual disabilities: inequalities and opportunities. Future Internet, 5, 376-397. doi:10.3390/fi5030376

[3] Dannenberg P, Fuchs M, Riedler T, Wiedemann C. (2020). Digital transition by COVID-19 pandemic? The German food online retail. Journal of Economy and Human Geography, 111(3), 543-560. https://doi.org/10.1111/tesg.12453

[4] Lyengar K, Upadhyaya GK, Vaishya R, Jain V. (2020). COVID-19 and applications of smartphone technology in the current pandemic. Diabetes & Metabolic Syndrome: Clinical Research and Reviews, 14(5), 733-737. https://doi.org/10.1016/j.dsx.2020.05.033

[5] De R, Pandey N, Pal A. (2020). Impact of digital surge during Covid-19 pandemic: A viewpoint on research and practice. International Journal of Information Management, 55. https://doi.org/10.1016/j.ijinfomgt.2020.102171

[6] Petrie HL, Savva A, Power C. (2015). Towards a unified definition of web accessibility. ResearchGate. doi:10.1145/2745555.2746653

[7] Courtenay K, Perera B. (2020). COVID-19 and people with intellectual disability impacts of a pandemic. Irish Journal of Psychological Medicine, 37 (3), 231-236. doi:10.1017/ipm.2020.45

[8] Cinquin PA, Guitton P, Sauzéon H. (2020). Designing accessible MOOCs to expand educational opportunities for persons with cognitive impairments. Behaviour & Information Technology, 40(10), 1101-1119. https://doi.org/10.1080/0144929X.2020.1742381

[9] Menta O, Androutsou T, Anastasiou A, Koumpouros Y, Matsopoulos G, Koutsouris D. (2020). A three-module proposed solution to improve cognitive and social skills of students with attention deficit disorder (ADD) and high functioning autism (HFA): innovative technological advancements for students with neurodevelopmental disorder. Association for Computing Machinery, 72, 1-7. https://doi.org/10.1145/3389189.3398003

[10] Giraud S, Thérouanne P, Steiner DD. (2018). Web accessibility: Filtering redundant and irrelevant information improves website usability for blind users. International Journal of Human-Computer Studies, 111, 23-35. https://doi.org/10.1016/j.ijhcs.2017.10.011

[11] Dove E, Astell A. (2017). The use of motion-based technology for people living with dementia or mild cognitive impairment: A literature review. Journal of Medical Internet Research, 19 (1). doi: 10.2196/jmir.6518

[12] Johansson S, Gulliksen J, Gustavsson C. (2021). Disability digital divide: the use of the internet, smartphones, computers and tablets among people with disabilities in Sweden. Universal Access in the Information Society, 20, 105–120. https://doi.org/10.1007/s10209-020-00714-x

[13] Alajarmeh N. (2021). The extent of mobile accessibility coverage in WCAG 2.1: sufficiency of success criteria and appropriateness of relevant conformance levels pertaining to accessibility problems encountered by users who are visually impaired. Universal Access in the Information Society. https://doi.org/10.1007/s10209-020-00785-w

[14] Jones P. (2019). Contexts of co-creation: Designing with system stakeholders. Systemic Design, Translational Systems Sciences, 8, 3-52. https://doi.org/10.1007/978-4-431-55639-8_1

[15] Schuurman D, Coorevits L, Logghe S, Vandenbroucke K. (2015). Co-creation in living labs: exploring the role of user characteristics on innovation contribution. International Journal of Services Sciences 5(3-4),199-219. doi:10.1504/IJSSCI.2015.074218

Challenges of Trustable AI and Added-Value on Health
B. Séroussi et al. (Eds.)

doi:10.3233/SHTI220544

iCompanion: A Serious Games App for the Management of Frailty

Antonios SYKOUTRIS, Angelina KOUROUBALI, Dimitrios G. KATEHAKIS and Haridimos KONDYLAKIS
FORTH-ICS, Heraklion, Crete, Greece

Abstract. The term frailty is often used to describe a particular state of health, related to the ageing process, often experienced by older people. The most common indicators of frailty are weakness, fatigue, weight loss, low physical activity, poor balance, low gait speed, visual impairment and cognitive impairment. The objective of this work is the creation of a serious games mobile application to conduct elderly frailty assessments in an accurate and objective way using mobile phone capabilities. The proposed app includes three games (memory card, endless runner, and clicker) and three questionnaires, aiming towards the prediction of signs of memory and reflection deterioration, as well as endurance and strength. The games, when combined with a set of qualified questionnaires, can provide an efficient tool to support adults in identifying frailty symptoms and in some cases prevent further deterioration. At the same time the app can support older adults in improving physical and mental fitness, while gathering useful information about frailty.

Keywords. Frailty, mHealth, questionnaires, digital health, serious games

1. Introduction

Frailty is a very common clinical condition for the elderly [1], [2]. It leads to the weakening of various functional areas, such as physical frailty which includes unintentional weight loss, self-reported exhaustion, weakness, slow walking speed and low physical activity; cognitive frailty which is a distinct clinical concept of simultaneous physical frailty and cognitive impairment in the absence of concurrent dementia; psychological frailty such as loss of resilience in cognitive, mood and motivational components; and social frailty which includes loss of social resources and behaviors that are important for an individual's social needs [3].

For the past few years, personal health apps [4] [5]and android/ios applications have exploded in popularity reestablishing and even reinventing the way people, connect, exchange information, interact and socialize. In recent years, various studies have been conducted considering the use of new devices that allow for the detection and monitoring of frailty without expensive facilities [6]. There is a great need to create applications specifically designed for pre-frail and frail older adults that focus on their unique needs and goals, as such software is currently not common in the market [7].

The aim of this work is to create a mobile application that provides a user-friendly interface to engage older adults in a comfortable way to spend time on serious games and other activities that could support the detection and improvement of frailty symptoms. Evidenced based questionnaires are incorporated in the software for the collection of older adult data that could assist healthcare professional in identifying signs of frailty

and assess the effectiveness of the treatment. iCompanion is a prototype that could provide the basis for a functional digital health app to be used in real-world scenarios.

The rest of the paper is structured as follows: In Section 2 we present an overview of the methods used for developing the app and then in Section 3 we present the serious games app that allows older adults to play games and answer daily questionnaires. The application, facilitates the interaction between older adults and healthcare professionals as it logs the progress of users engaging with the serious games. Also, it provides and overview of the answers to the questionnaires over time. A dashboard shows the effectiveness of the user interactions with the app in a monthly graph. Finally, Section 4 concludes this paper and presents directions for our next steps.

2. Methods

The focus of this work is the design and development of iCompanion, a digital health app prototype that uses SG to manage and support frailty in older adults.

Table 1. SG examples

Virtual Supermarket game[1]	The user must pick a list of items from the supermarket. It contains only a short shopping list to train the short memory.	This Game is used for diagnostic purposes, physical and cognitive status of older persons.
Frail Red Wings[2]	Flying a plane using a dynamometer in hand.	The user controls the plane with the dynamometer and applies pressure to avoid obstacles. The game targets grip strength and helps the user strengthen it.
FRED[3]	Exercise game that uses biofeedback to ensure that the game is performed under cardio-healthy conditions.	The game supports the users in finding meaning in the activity. Each movement is designed by considering both biomechanical and neuromotor parameters and evidence features of sufficient extent to be recognised by the Kinect sensor.

The methodology used included initially a review of selected serious games in the field to identify the most important elements that are important to be included in SG applications. In the last decades there has been a growing interest in employing Information and Communication Technologies (ICT) to facilitate the assessment of functional impairments of older adults and support them in their everyday activities [8] [9][10]. In addition to being important for health assessment, ICT can also play a key role in the treatment, stimulation, and rehabilitation of health conditions. Serious Games (SG) are digital applications specialized for purposes other than entertaining, such as education and information as well as enhancement of user aptitudes and/or cognitive and physical functions. SG have also been shown to facilitate improvement of different users' functional and cognitive traits in relation to a healthcare treatment plan. SG provide the opportunity for healthcare providers to monitor the player, while tracking game data that can be used for subsequent analysis and comparison [11]. Several SG applications for frailty have been developed that provide different features for engaging the elderly. FrailSafe, a Horizon 2020 research and innovation programme (https://frailsafe-project.eu/), has developed several mini-games for frailty detection and prevention. Within the suite of SG are memory games, reflexive games that are developed to evaluate and stimulate the visual reflex and motor speed of people afflicted by frailty and sound

[1] https://frailsafe-project.eu/news/59-supermarket
[2] https://frailsafe-project.eu/news/61-red-wings
[3] https://addi.ehu.es/handle/10810/22749

sequence games that use sound to create sequences that the patient can identify and replicate. Some SG examples are presented in Table 1.

However, in an ageing society, it is necessary to establish new alternatives that focus on the needs of older adults while increasing their perceived quality of life, combining SGs and established clinically validated questionnaires that are daily used in clinical practice [12]. Based on the survey, a list of features was developed that was then explored during in depth interview with healthcare providers. Overall, two interviews were conducted with three clinicians. During the interviews the list of features was presented to the clinicians that were then rated to identify the most useful ones to be included in a series of mini-games. In addition, the clinicians identified the most relevant questionnaires to be used for identifying prefrailty and frailty.

3. Results

Consolidating all requirements identified, the iCompanion mobile app was build using a modular architecture depicted in Figure 1. The mobile app includes the user management module, three mini-games for frailty assessment and three standard clinically validated questionnaires for monitoring citizen status. All data are stored in a secure repository online whereas users are authenticated using national infrastructures. In the sequel we describe in detail the various modules of the Mobile app.

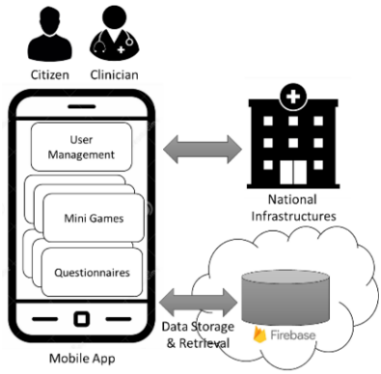


Figure 1. The High-level architecture of the mobile app.

User Management. The users of our application can be either clinicians or citizens. They can all select their role upon registration and be authenticated through the relevant national infrastructures. Currently the mobile app is connected to the Greek national infrastructures but can be easily extended to support other authentication services (e.g. Facebook, Google etc.). Citizens can track their own progress and play the mini-games which can then bring to the clinicians for discussing the graphs and their progress. For clinicians the mobile app allows adding supervised citizens, that can use the application while in the waiting room to check their progress. In addition, the clinician account has the option to complete the two STRATIFY questionnaires, whereas the citizens have the opportunity to complete the EQ-5D-5L self-reported questionnaire. All results are logged and the computed scales can be discussed with the clinicians. Further, the iCompanion mobile app includes three serious mini-games for assessing citizen status using a variety of tools:

Memory card game. Memory card games are games in which all cards are laid face down and each turn the player(user) flips two cards of his/her choosing trying to identify similar ones. The time in which the user finishes the game is very important because it stands as a good measurement of how fast the user can remember the correct pairs and how developed is a short memory. In our variance of the classic card game, the player has to identify the pair that is not identical, but where one card is similar to another card - for example, one card could be a monkey and the other card could be a banana. The reason behind this particular approach is that you tend to exercise the short memory a bit

more by forcing the user to correlate the two different items. Each time the user plays the score is saved and at the end compared with other players (anonymously) along with his own past performance. The user can gradually move to harder versions (10 pairs, 20 pairs etc.). User performance in similar games has been an indicator of Mild Cognitive Impairment (MSI) in the past [13], so by evaluating the results over time the clinicians can monitor the citizens memory performance and to identify signs of deterioration.

Endless runner game. Endless runner is a SG where the player character is forced to run for an infinite amount of time while dodging obstacles. Such SGs are proven to train the reflexes of the user, by procedurally generating more obstacles [12]. The gyroscope sensor of the mobile is responsible for the autorotation of the screen and view on the screen whenever a phone is rotated. Combining the endless runner subgenre with the gyroscope movement, in the implemented SG a ball moves in a linear path trying to avoid obstacles and can continuously assess and train the reflexes of the patient. Additionally, by collecting gold coins found within the game, the patient gains score points. The highest score is being saved by the software to keep track as a reflection measurement providing a mean to monitor reflection skills.

Clicker game. Clicker games are video games whose gameplay consists of the user performing simple actions such as tapping on the screen repeatedly. The reason behind making such a game is that elderly people with advanced forms of frailty are shown to lack grip endurance [13]. This is important considering that clicker games test grip endurance and strength. In our version, there is a sword stuck in a rock waiting to be lifted. The user must continuously press the pull button to reveal the sword. There is a countdown of 60 seconds that the task should be finished before resetting the sword. The more grip endurance the patient has the faster will be able to pull the sword. All scores are saved and can help the clinicians monitor grip endurance.

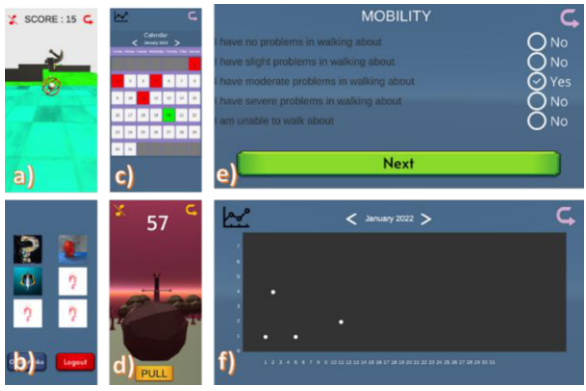


Figure 2. The GUI of the three mini-games, i.e. a) the endless run game, b) the memory game, c) the endless run, d) the mini calendar, e) a questionnaire and f) some statistics.

Questionnaires. Three clinically validates questionnaires have been implemented for risk assessment. The answers each user provides are analyzed and risk scores are calculated in each case based on the provided answers. After completing a questionnaire, the users can review the relevant scores and also view a monthly graph. The first two questionnaires are completed by nurses/clinicians based on their interviews with the citizens. STRATIFY [15] questionnaires, have been developed to predict patients at high risk of falling with clinically useful sensitivity and specificity. It is increasingly used routinely in elderly care departments for this purpose. The citizens on the other hand can complete the EQ-5D-5L, a self-assessed, health related, quality of life questionnaire. The scale measures quality of life on a 5-component scale including mobility, self-care, usual activities, pain/discomfort, and anxiety/depression. As deteriorating quality of life has been highly related with frailty [15] the questionnaire provides a useful resource for monitoring citizens.

GUI. The GUI of the app is shown in Figure 2. All modules have been implemented using Unity, which enables delivering high-quality mini-games across all mobile devices.

4. Conclusion

Work conducted has incorporated gaming factors to enhance motivation towards frailty self-assessment. iCompanion is a low-cost tool that has the potential to support generalized frailty management without the need for expensive equipment. Additionally, the software resembles and functions as a medical software that can help patients with frailty prevent further deterioration. Moreover, the android/ios application supports a user-friendly, simple, and clean user interface that does not overwhelm the users. In future versions, users will have the ability to play a varied suite of serious games with more advanced features focused on frailty needs. Currently the app is loaded on tablets at the clinics. While patients are in the waiting room a nurse completes the questionnaires and gives the tablet to the patients for playing the mini-games. The results are collected and as soon as the patient is with the doctor, the doctor has in front of him a dashboard with the patient's progress over time in order to help him understand patient's status. Additionally, there are plans to expand and enhance the UI of the software. Finally, a detailed evaluation by both clinicians and citizens is already planned to follow, whereas a market research is being conducted to facilitate the development of a viable business.

References

[1] Clegg A, et al. Frailty in elderly people. The lancet. 2013;381(9868):752-62.
[2] Chen X, Mao G, Leng SX. Frailty syndrome: an overview. Clinical interventions in aging. 2014;9:433.
[3] Pilotto A, et al., A multidimensional approach to frailty in older people, Ageing research reviews 60 (2020).
[4] Katehakis DG, et al. Integrated care solutions for the citizen: personal health record functional models to support interoperability. EJBI 2017:13(1), 41-56.
[5] Maniadi E, et al. Designing a digital patient avatar in the context of the MyHealthAvatar project initiative. In 13th IEEE international conference on BioInformatics and BioEngineering 2013:1-4/
[6] Anabitarte-García F, et al. Early diagnosis of frailty: Technological and non-intrusive devices for clinical detection. Ageing Research Reviews. 2021;70:101399.
[7] Kouroubali A, Kondylakis H, Logothetidis F, Katehakis DG. Developing an AI-Enabled Integrated Care Platform for Frailty. In Healthcare 2022;10(3):443.
[8] Robert P, et al. Recommendations for the use of Serious Games in people with Alzheimer's Disease, related disorders and frailty. *Frontiers in aging neuroscience*, *6*,
[9] Smith J, Forster A, Young J. Use of the 'STRATIFY'falls risk assessment in patients recovering from acute stroke. Age and Ageing 2006;35(2):138-143.
[10] Kondylakis H, et al. Semantically-enabled Personal Medical Information Recommender.ISWC 2015.
[11] Shapoval S, et al. Biofeedback Applied to Interactive Serious Games to Monitor Frailty in an Elderly Population. *Applied Sciences*, 2021;11(8):3502.
[12] Mugueta-Aguinaga I, Garcia-Zapirain B. Is technology present in frailty? Technology a back-up tool for dealing with frailty in the elderly: a systematic review. *Aging and disease* 2017;*8*(2):176.
[13] Sood, Pallavi, et al. Nonimmersive brain gaming for older adults with cognitive impairment: A scoping review. The Gerontologist 2019;59.6:e764-e781.
[14] Herdman M, et al. Development and preliminary testing of the new five-level version of EQ-5D (EQ-5D-5L). Quality of life research 2010;20(10):1727-1736.
[15] Langlois F, et al. The multiple dimensions of frailty: physical capacity, cognition, and quality of life. International Psychogeriatrics 2012;24(9):1429-1436.

Challenges of Trustable AI and Added-Value on Health
B. Séroussi et al. (Eds.)

doi:10.3233/SHTI220545

Classification of Digital Mental Health Interventions: A Rapid Review and Framework Proposal

Marie-Pierre GAGNON[a,b1] , Maxime SASSEVILLE[a,b] and Annie LEBLANC[b,c]
[a] *Faculty of Nursing Sciences, Université Laval, Québec, Canada*
[b] *VITAM Centre de recherche en santé durable*
[c] *Department of Family and Emergency Medicine, Faculty of Medicine, Université Laval, Québec, Canada*

Abstract. The modern context of mental health interventions asks for an inclusion of digital solutions to the face-to-face approach, providing better access and reduced inequity for patients. The current classification of digital mental health interventions can be system specific (mobile apps) or general (virtual therapy), which causes inadequacy in applications. The goal of this study was to develop a framework to improve digital mental health interventions classification. We performed a rapid review of the literature on existing digital mental health interventions frameworks. We identified four existing frameworks, extracted their purpose, categories and items, completed a thematic analysis and formulated a four domains framework proposal. This framework allows to classify digital mental health interventions on their system, function, time and facilitation, which should facilitate our understanding of the effect of singular characteristics on patient outcomes.

Keywords. Digital health, digital interventions, mental health, classification

1. Introduction

Technologies in healthcare have received an unpreceded interest amidst the COVID-19 pandemic. Many people living with a chronic physical or mental health condition experienced follow-up disruption with care providers during the pandemic [1]. People living with a chronic disease are at a higher risk of developing mental health problems [2]. In a pandemic context affecting particularly the elderly and chronically ill people, combined with public health measures that limit social interactions, mental health status is particularly affected [3].

In Canada, common chronic conditions are mostly managed in primary care by care teams including physicians, nurses, social workers and psychologists. While some services could be maintained at distance during the pandemic, mostly with personalized digital solutions, in many instances neither patients nor providers were prepared for this digital shift. Most clinicians had never provided digital health services before, and healthcare managers and decision makers did not know which digital health interventions

[1] Corresponding Author: Marie-Pierre Gagnon, Faculty of Nursing Sciences, Université Laval, 1050, avenue de la Médecine, Québec, Canada, G1V 0A6; E-mail: marie-pierre.gagnon@fsi.ulaval.ca

should be implemented. Although digital health was seen as an opportunity to maintain care continuity for people living with chronic conditions, the lack of knowledge about digital health solutions, their safety and effectiveness limited their implementation.

A knowledge synthesis was conducted to gather evidence regarding safe and effective digital mental health interventions (DMHI) for people living with a chronic condition [4]. This project was funded by the Canadian Institutes of Health Research (CIHR), and involved a team of researchers with expertise in content and methods, and knowledge users (clinicians, managers and patients). To provide knowledge users with effectiveness evidence for different types of DMHI, we were faced with the lack of consensus and comprehensiveness of current classification of DMHI.

The aim of this paper is to present the development of a classification framework for DMHI to provide answers to knowledge users on digital solutions that could be implemented for the prevention, detection and management of mental health conditions in primary care. Given the lack of a comprehensive classification framework, our systematic approach provides a basis to organize knowledge on DMHI and compare different interventions.

2. Methods

This parallel methodological study was conducted in the context of a larger knowledge synthesis of digital mental health interventions in primary care [4]. We performed a rapid review, applying the process and methods of a systematic review in a streamlined and accelerated way [5]. We limited the search to two main databases, MEDLINE and PsycInfo with the following strategy: ("mental health" OR "anxiety" OR "depression") AND ("telemedicine" OR "eHealth" OR "mHealth" OR "telehealth") AND ("classification" OR "taxonomy" OR "framework"). We included any peer-reviewed research paper describing a classification of any sort for digital mental health intervention, published in the last 10 years in French or in English. Publications were excluded if they did not describe their classification method or if they targeted a specific domain other than mental health. One researcher screened the title and abstract and the full-text was appraised by two researchers and discussed before inclusion. Data from the classifications identified were extracted. We used a thematic analysis to do a theme reduction and structure [6]. Second, we identified classification themes using thematic content analysis, and provided their definitions and attributes. Third, we conducted a targeted literature search of definitions of identified digital technologies to provide a common understanding of the terms used in the classification system. Fourth, the framework was validated through research team discussions. Finally, we applied the proposed classification framework to the DMHI found in the systematic review.

3. Results

The search strategy yielded 671 results in the database, from which 146 were duplicates. An additional record was included from the WHO website. We completed a full-text evaluation of 16 papers, from which 4 were retained (Figure 1).

3.1. Existing Classifications

The WHO Digital Health Intervention (DHI) framework proposes system categories that represent the types of technological applications and information systems designed to deliver one or more digital health interventions [7]. Liverpool et al. conduct a review of

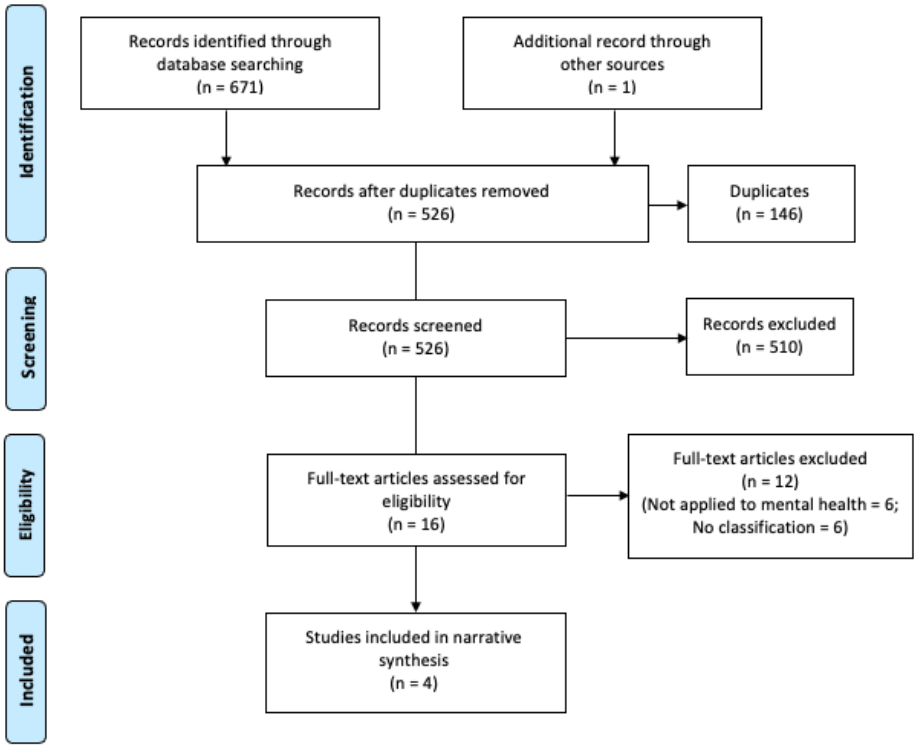


Figure 1. PRISMA flow diagram.

the literature to identify modes of delivery and goals of DMHI. They used a classification based on six modes of delivery and seven goals [8]. Grajales et al. offer a description of ten types of services in relation with social media [9]. The goal of this study was to conduct a narrative review of the literature to present how social media are being use in digital mental health interventions. Chan et al. described seven goals of mobile apps [10]. The purpose of this article was to present a framework to adequately evaluate mobile applications used in mental health. This framework is based solely on experts' opinions, but considers current policies and guidelines.

There was a total of 55 categories proposed in the four retained frameworks. Most of the frameworks classified items relying on one level category, except for Liverpool et al. that detailed modes of delivery apart from goals of the intervention. The frameworks from Liverpool et al. and the WHO are general purpose frameworks. For their part, Grajales et al. detailed subtypes of social media, and Chan et al. detailed functions of mobile applications. For all identified frameworks, they lacked important aspects related

to mental health interventions, such as the role of the therapist and whether the intervention was synchronous or asynchronous.

From the available evidence reviewed, we propose a four-dimensional framework. It comprises the notions of System, Function, Time and Facilitation (Table 1). The proposed classification allowed for targeted analysis to gather information on unique features (use of prompts and alerts) or general criteria (self-administered interventions). However, some interventions could fit in more than one category, pointing to the need for a classification system that could adapt to multifaceted interventions.

Table 1. Proposed classification framework of digital mental health interventions

System	Function (*sub-function*)	Time	Facilitation
1. Internet or Website 2. Computer (software) 3. Mobile app 4. Electronic messaging (email, SMS) 5. Electronic health record 6. Telehealth (telemedicine, telepsychiatry) 7. Virtual reality/ augmented reality 8. Robot 9. Connected devices 10. Social media 11. Other system	A. Decision support *a) Screening* *b) Prompts and alerts* B. Communication *a. Transmission of information (one way)* *b. Communication (with healthcare provider)* *c. Communication (peer to peer, e.g., virtual peer group for clients)* C. Therapy *a. Cognitive Behavioural Therapy (CBT)* *b. Other psychotherapy* *c. Gamification* D. Monitoring a. *Provider monitoring* *b. Self-monitoring* E. Other function	=. Synchronous +. Asynchronous	&. Entirely supported by healthcare providers @. Partially supported by healthcare providers #. Self-administered

4. Discussion

We developed a classification framework for DMHI for digital mental health solutions that could be implemented in primary care. This work aimed to answer a specific need of knowledge users about which digital interventions could be used to ensure the prevention, detection, or management of common mental health issues in people living with a chronic condition. Limits of current classifications were highlighted, and a new framework was developed to allow knowledge users access more specific evidence about DMHI that could be implemented.

This project was conducted in a very short turnaround, and we limited the number of databases consulted to only the most relevant, while following rapid review guidelines. Thus, it is likely that other relevant classification frameworks exist but were not considered.

5. Conclusion

A classification framework for digital mental health interventions was developed through a rapid review. We propose four dimensions (System, Function, Time and Facilitation) that can be used to describe and compare digital interventions aimed at the prevention, detection, or management of common mental health issues.

References

[1] Marshall EG, Breton M, Cossette B, Isenor JE, Mathews M, Ayn C, et al. The PUPPY Study–Protocol for a Longitudinal Mixed Methods Study Exploring Problems Coordinating and Accessing Primary Care for Attached and Unattached Patients Exacerbated During the COVID-19 Pandemic Year. medRxiv. 2021.

[2] Watson LC, Amick HR, Gaynes BN, Brownley KA, Thaker S, Viswanathan M, et al. Practice-based interventions addressing concomitant depression and chronic medical conditions in the primary care setting: a systematic review and meta-analysis. Journal of primary care & community health. 2013;4(4):294-306.

[3] Sepúlveda-Loyola W, Rodríguez-Sánchez I, Pérez-Rodríguez P, Ganz F, Torralba R, Oliveira D, et al. Impact of social isolation due to COVID-19 on health in older people: Mental and physical effects and recommendations. The journal of nutrition, health & aging. 2020:1-10.

[4] Sasseville M, LeBlanc A, Boucher M, Dugas M, Mbemba G, Tchuente J, et al. Digital health interventions for the management of mental health in people with chronic diseases: a rapid review. BMJ Open. 2021.

[5] Garritty C, Gartlehner G, Nussbaumer-Streit B, King VJ, Hamel C, Kamel C, et al. Cochrane Rapid Reviews Methods Group offers evidence-informed guidance to conduct rapid reviews. Journal of Clinical Epidemiology. 2020.

[6] Ouzzani M, Hammady H, Fedorowicz Z, Elmagarmid A. Rayyan—a web and mobile app for systematic reviews. Systematic reviews. 2016;5(1):1-10.

[7] Miles MB, Huberman AM, Saldaña J. Qualitative data analysis: A methods sourcebook: Sage publications; 2018.

[8] World Health Organization. Classification of digital health interventions v1. 0: a shared language to describe the uses of digital technology for health. World Health Organization; 2018.

[9] Liverpool S, Mota CP, Sales CM, Čuš A, Carletto S, Hancheva C, et al. Engaging children and young people in digital mental health interventions: systematic review of modes of delivery, facilitators, and barriers. JMIR. 2020;(6):e16317.

[10] Grajales FJ III, Sheps S, Ho K, Novak-Lauscher H, Eysenbach G. Social media: a review and tutorial of applications in medicine and health care. JMIR. 2014;16(2):e13

Challenges of Trustable AI and Added-Value on Health
B. Séroussi et al. (Eds.)

doi:10.3233/SHTI220546

Impact of the Text Simplification on Understanding

Oksana IVCHENKO [a] and Natalia GRABAR [a]

[a] *CNRS, Univ. Lille, UMR 8163 - STL - Savoirs Textes Langage, F-59000 Lille, France*

Abstract. The reduction of the linguistic complexity of medical texts to make them more understandable to a larger population is an important task. The simplification of texts involves several steps, among which our study focuses on the definition of complex constructions and on study of the impact of the simplification. For this study, we selected 20 texts from the medical domain on different topics, namely drugs, diseases, substances, and medical institutions. We identified complex linguistic constructions and carried out their manual simplification at syntactic, lexical and semantic levels. We then designed a questionnaire to test comprehension of the texts and conducted a study with 26 participants. The results of this study shows that simplified texts obtained higher number of correct answers than technical texts. This difference is statistically significant. The self-evaluation questionnaire, done at the beginning of the test, indicates that the participants tend to overestimate their understanding of medical information. Besides, there is no correlation between the time taken to complete the interview and the correct answers provided.

Keywords. Simplification, Readability, Evaluation, Correlation, France

1. Introduction

The purpose of text simplification is to reduce the linguistic complexity of the content and to make it more understandable for a given population. In the case of medical texts, non-specialist speakers are not familiar with technical terms and expressions, which makes it very difficult to read and understand these texts [1]. Thus, even though medical literature is becoming more freely available online, these documents generally require a level of understanding beyond the capabilities of the average reader [2]. This can lead to a lack of understanding of medical information by patients. Indeed, medicine is currently one of the most rapidly evolving branches of science. There is consequently a growing need for automatic methods to make medical texts more accessible [3]. Yet, before creating automatic simplification systems, it may be useful to know what is the real impact of simplification on comprehension of medical texts.

The simplification of texts usually involves several aspects (detection of complex sequences, lexical and syntactic simplification, evaluation...). The objective of our work is to study where lies the complexity of medical texts and to observe the impact of the text simplification on their understanding. We propose a qualitative study based on interviews with non-specialist speakers.

In what follows, we first introduce the material used and the methods. We then present the results obtained and we discuss them. Finally, we conclude with some issues for future work.

2. Material and Methods

2.1. Source Corpus and Creation of Working Data

The texts studied are randomly selected from the CLEAR corpus [4] and its three genres (encyclopaedic articles, drug leaflets and Cochrane scientific summaries). These texts address diseases (leukaemia, anaemia), drugs (Tracarium), substances (antigens, anti-cholinergics), and institutional facilities. From these, we selected 20 segments with 1 to 3 sentences. Then, we manually identified the complex terms and constructions, and simplified them manually according to three linguistic levels: lexical, syntactic, and semantic.

As has been noticed, lexical complexity of terms has an important impact and, even in short sentences, may prevent the understanding of the whole sentence [5]. Contrary to the general language, short medical terms (such as *ectasia* or *anthrax*) do not make the text more understandable or informative [6]. Hence, we exploit the frequency of words and terms as an indicator of their complexity, as frequent words are read more often. The frequency is provided by *Lexique.org* [7]. Unfrequent and rare terms are replaced by their synonyms or explanations, like *gériatrique* (*geriatric*) replaced by *les personnes âgées* (*[for] elderly people*). These are searched in the *bio-top* resource[1], which is created by medical specialists and qualified professionals. Similarly, the abbreviations are developed using this resource as well, like for *la molécule AMPc* (*the cAMP molecule*) meaning *adénosine monophosphate cyclique* (*cyclic adenosine monophosphate*).

Syntactic simplification involves reducing the grammatical complexity of a text while preserving its information content and meaning [8]. In our sample, we observe that the texts are generally long, with complex subordinate and coordinated sentences. It is therefore necessary to rewrite them into shorter sentences to improve their readability [9]. The syntactic rules implemented can be divided into several types: (1) segmentation of subordinate, coordinated and conjunctive clauses; (2) rewriting of sentences according to a simple word order (Subject-Verb-Complement); (3) transformation of passive sentences into active sentences; (4) modification of negative sentences into positive sentences; (5) preference for the present tense of verbs.

At the semantic level, the most important issue is that the texts remain coherent and clear, and that the information can be understood in its context [10]. We make three additional types of transformations: (1) reorganization of sentences for a better presentation of information, (2) deletion of secondary segments that do not affect the general meaning of the text, (3) addition of examples or explanations for a better understanding.

Table 1. Number of words and sentences: minimal, maximal and average value

Indicator	*Technical texts*		*Simplified texts*	
	words	sentences	words	sentences
min	7	1	16	1
max	78	3	75	4
average	33.85	1.25	38.85	2.15

Table 1 presents the results of the simplification. We can see that simplified texts become longer: they contain more sentences and words. Indeed, the simplification of medical texts often requires addition of information.

2.2. Creation of the Questionnaire and the Interviews

To evaluate the comprehension of technical and simplified texts, we create a questionnaire with multiple choice questions (MCQs). We use different types of questions: definitional, factual, with requests for precision or description and thus we formulated 40 questions. The comprehension of each segment is addressed with two questions: one question on the beginning of the text and one on the end of the text. Four responses are proposed for each question in random order (one correct answer, two wrong answers (distractors), and *I don't know*).

We did a pre-test of the questionnaire with two participants. Following the pre-test, we did not notice any ambiguities in the questions or in the answers. However, we modified three questions to have a greater variety in the types of questions.

At the beginning of the interviews, the participants had to complete a self-evaluation test HLS-EU16 that focuses on the overall understanding of medical information [11]. In this self-evaluation form, the participant indicates, on a scale going from very easy (1) to very difficult (5), how easy it is for him/her to understand medicine-related information. For example, how well the participant understands what the doctor tells him/her, what the results of laboratory tests show, etc. Then, the participants had to answer the main questionnaire after reading each text. The answers collected are analyzed statistically.

3. Results and Discussion

A total of 26 volunteers took part in the study. Their native language was French and they lived in different parts of France as well as abroad. They had no medical education. The correct answers collected for all segments are presented in Figure 1. We obtain on average 16.8 (median 18.5) correct answers for the technical texts (in red), and 22.3 (median 23) for the simplified texts (in blue). The participants spent an average of 14.19 minutes answering the questionnaire.

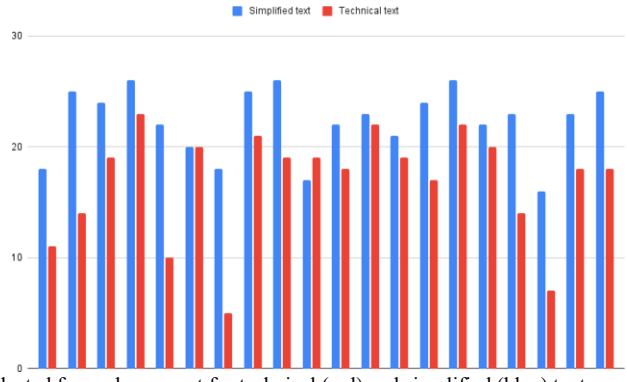


Figure 1. Correct answers collected for each segment for technical (red) and simplified (blue) texts

To test the hypothesis H0 that the observed random variable follows the theoretical distribution law, we conducted a Chi2 test. The analysis of the responses with the Chi2 test shows that there is a statistical significance in the comprehension between the two versions of the texts with $p < 0.00001$. Therefore, this indicates that simplified texts are indeed easier to understand. The analysis for each pair of technical and simplified texts by Student's t-test shows that $p = 0.00009$.

This value is statistically significant and also indicates that the medical texts are easier to understand after simplification. However, if we look at the pairs of texts individually, 6 out of 20 pairs show the p value >0.05. Hence, the p is not statistically significant here, which indicates that some simplified versions of texts are not easier to understand than their original versions. This may be due to the complexity of information, even after simplification, or to the fact that information is added during the simplification, which may lead to sentences syntactically more complex.

Figure 2 shows the Pearson correlation test between the self-evaluation and the number of correct answers. The r-value is -0.4654, which indicates that participants tend to overestimate their ability to understand medical information in everyday life.

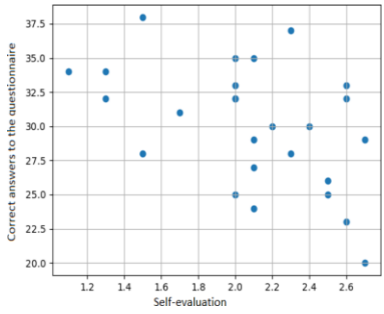


Figure 2. Correlation between self-evaluation and correct answers from participants

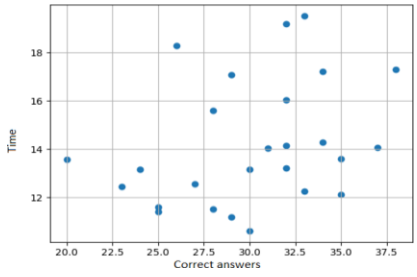


Figure 3. Correlation between correct answers and the time required for completion

Figure 3 shows a weak correlation ($r = 0.3053$) between the number of correct answers and the time taken to answer the questionnaire. At the same time, the p-value (0.12) does not confirm the statistical significance of this analysis. Therefore, we cannot say that participants who spent more time reading the texts and answering questions have a higher probability of providing correct answers.

We considered all the modifications at once (lexical, syntactic and semantic) and, for this reason, cannot indicate which simplification level is the most salient. Yet, we can state that modifications at several levels are required.

4. Conclusion and Perspectives

Our study shows that medical texts become more comprehensible after simplification, and this result is statistically significant. We have manually simplified the texts on three linguistic levels, and each text tended to be simplified on several levels at once. We performed manual simplification because the aim of the experiment was to determine which features in medical texts to consider when simplifying in order to optimize further work on automatic simplification. So we cannot say exactly which linguistic constructions (syntactic, lexical or semantic) make the text difficult to understand. Yet, we can state that the simplification should cover several levels of simplification. The results of the self-evaluation of comprehension of medical texts show that people tend to overestimate their knowledge. This prompts a search for more objective methods of identifying difficult segments in technical texts. For instance, we can use eye-tracking techniques paying particular attention to the use of large amounts of data and a more important number of participants.

Acknowledgements. This work was partly funded by the French National Agency for Research (ANR) as part of the *CLEAR* project (*Communication*, *Literacy*, *Education*, *Accessibility*, *Readability*), ANR-17-CE19-0016-01.

References

[1] Paetzold GH, Specia L. A Survey on Lexical Simplification. J Artif Int Res. 2017 Sep;60(1):549–593.
[2] https://doi.org/10.1613/jair.5526
[3] Kloehn N, Leroy G, Kauchak D, Gu Y, Colina S, Yuan NP, et al. Improving Consumer Understanding of Medical Text: Development and Validation of a New SubSimplify Algorithm to Automatically Generate Term Explanations in English and Spanish. J Med Internet Res. 2018.https://www.jmir.org/2018/8/e10779/
[4] Institute of Medicine. Health Literacy: A Prescription to End Confusion. Nielsen- Bohlman L, Panzer AM, Kindig DA, editors. Washington, DC: The National Academies Press; 2004.Availablefrom:https://www.nap.edu/catalog/10883/ health-literacy-a-prescription-to-end-confusion.
[5] Grabar N, Cardon R. CLEAR – Simple Corpus for Medical French. In: Workshop on Automatic Text Adaption (ATA); 2018. p. 1-11.
[6] Collet T. Obstacles lexico-sémantiques à la lecture réussie d'un texte de spécialité. TTR. 2014;27(1):123-48. https://doi.org/10.7202/1037121arGu Y, Leroy G, Kauchak D. When synonyms are not enough: Optimal parenthetical insertion for text simplification. AMIA Annual Symposium proceedings AMIA Symposium. 2017;2017:810-9.
[7] https://pubmed.ncbi.nlm.nih.gov/29854147/
[8] New B, Pallier C, Brysbaert M, Ferrand L. Lexique 2 : A new French lexical database. Behavior Research Methods, Instruments & Computers. 2004;36(3):516-24.
[9] https://pubmed.ncbi.nlm.nih.gov/15641440/
[10] Siddharthan A. Syntactic simplification and Text Cohesion. No. 597 in Technical Reports. University of Cambridge; 2004.http://www.cl.cam.ac.uk/techreports/UCAM-CL-TR-597.pdf
[11] Xu W, Napoles C, Pavlick E, Chen Q, Callison-Burch C. Optimizing Statistical Machine Translation for Text Simplification. Transactions of the Association for Computational Linguistics. 2016;4:401-15.
[12] https://aclanthology.org/Q16-1029.pdf
[13] Brouwers L, Bernhard D, Ligozat AL, François T. Simplification syntaxique de phrases pour le français (Syntactic Simplification for French Sentences) [in French]. In: JEP-TALN-RECITAL; 2012.
[14] https://aclanthology.org/F12-2016
[15] Rouquette, T N, P L, den Broucke S et al V. Validity and measurement invariance across sex, age, and education level of the French short versions of the European Health Literacy Survey Questionnaire.; 2018. https://pubmed.ncbi.nlm.nih.gov/30521552/

Challenges of Trustable AI and Added-Value on Health
B. Séroussi et al. (Eds.)

doi:10.3233/SHTI220547

An Agile Approach to Accelerate Development and Adoption of Electronic Product Information Standards

Catherine CHRONAKI[a,1], Craig ANDERSON[b], Jens Kristian VILLADSEN[c], Petter HURLEN[d], Giorgio CANGIOLI[a], Giovanna Maria FERRARI[e] and Anne MOEN[f,g]
[a] *HL7 Europe, Brussels, Belgium*
[b] *Pfizer, Toronto, Canada*
[c] *Trifork, Copenhagen, Denmark*
[d] *Akershus University Hospital, Lørenskog, Norway*
[e] *Pfizer, London, UK*
[f] *Institute for health and society, University of Oslo, Oslo, Norway*
[g] *Norwegian Center for eHealth Research, Tromsø, Norway*

Abstract. The Medical Product Information found in most medication boxes offer a wealth of information, including terms of active ingredients, excipients, indications, dosage, route of administration, risks, and safety information. Digital health services that help patients, their care givers, and health professionals to manage medication, can be improved with tailored information based on user profile, the patient's Electronic Health Record (EHR) summary, and Medicinal Product Information. The electronic Product information (ePI) comprises the summary of product characteristics, package leaflet, and product label. The European Medicines Agency released in 2021 the first version of the EU proof-of-concept ePI standard based on HL7 FHIR. The Gravitate-Health project uses this common standard as a springboard to implement a federated open-source platform and services that helps advance access, understanding, and adherence by providing trusted medicinal information in an interoperable and scalable way. In this paper, we present the agile technical approach and co-creation process to design, test, and progressively mature interoperability working with the HL7 Vulcan Accelerator and FHIR connectathons.

Keywords. Citizen Empowerment, interoperability, electronic product information, International Patient Summary, HL7 FHIR, regulatory science

1. Introduction

In the European Union, a medicine's product information refers to the summary of product characteristics (SmPC) that is intended for healthcare professionals (HCPs), labelling (outer and inner packaging information) and package leaflet (PL) intended for for patients / consumers that is generally included as a printed copy in the medicines package [1]. The regulated and scientifically validated information included in SmPC assists HCPs in prescribing and dispensing the medicine. PLs included in the medicine

[1] Corresponding Author, Catherine Chronaki, HL7 Europe, 38-40 Square de Meeus, Brussels, 1000, Belgium; E-mail: chronaki@HL7europe.org

package inform patients and consumers about safe use. Prior to marketing authorization, a rigorous regulated process ensures that the paper PL can be read and understood by its users, patients and caregivers. However, in a digital world, the content of the paper PL is considered as a trusted yet underexploited information resource.

A report from the European Commission in March 2017, and a subsequent European Medicines Agency (EMA) action plan [2], identified areas where the SmPC and PL could be improved. Developing key principles for an ePI format was identified as the most pressing priority from a public health perspective and these key principles were finalized and published in January 2020 [1]. Further to this, the EMA ePI SetUP project was established to develop a common European format for ePI based HL7 FHIR® [3]. This ePI setup project started early in 2021, to create a proof-of-concept EU common standard specification in line with the key ePI principles. The draft specification was published in June 2021, together with an exploratory FHIR server that hosted sample data. A hands-on workshop was organized in July 2021, followed by a stakeholder consultation. Then, the common standard specification was approved by the EU Network Data Board [4], the next ePI implementation phase started. Despite the shift towards use of electronic formats, the primary form of delivery for Medicinal Product Information is still paper and regulatory processes built around this paradigm. While there are procedures and guidance on testing understanding of paper leaflets [3], such guidance is missing for ePI.

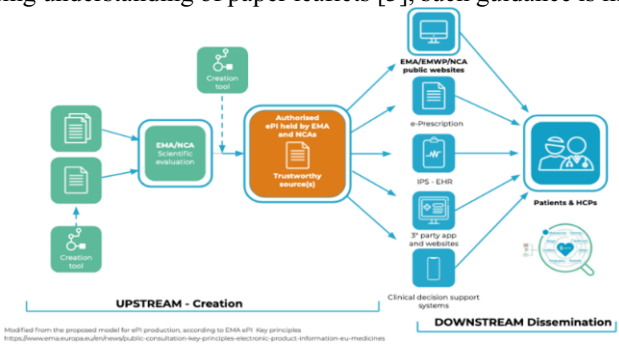


Figure 1. Gravitate-Health engages upstream, when creating and assessing ePIs and downstream, when tailoring ePI to user needs and preferences optimizing its impact on the understanding of the information.

The Gravitate-Health public-private partnership (www.gravitatehealth.eu) aims to offer citizens services that are tailored to their needs and help them access, understand and adhere to their medication therapy, supported by information conveyed by ePIs. To that end, the G-lens® combines users profile information with a summary of curated excerpt to their EHR, to highlight important ePI information. This approach will be tested in concrete user scenarios, starting in Europe and USA, and sharing experiences globally. The figure 1, above, describes the proposed data flow and process from the ePI creation (upstream) to ePI dissemination (downstream). Gravitate-Health aims to test the understanding of ePIs informing upstream. Furthermore, the G-lens® is envisaged to accelerate progress, in the 'Downstream dissemination' space, offering value to patients and their support network. Employing an HL7 FHIR standard-driven methodology, Gravitate-Health collaborates with the HL7 Vulcan accelerator's ePI project to incrementally advance and test solutions in HL7 FHIR connectathons (FCAT) three times a year. This methodology will be incorporated in the development of the Gravitate-Health HL7 FHIR IG (Implementation Guide). This methodology and preliminary findings are presented in the following sections. The discussion section reflects on the impact of the Gravitate-health approach and concludes with key points and next steps.

2. Methodology

Open standards and specifications are key to the vision of Gravitate-Health to improve access, understanding and adherence to medication treatment and exploit ePI along with information from EHRs, ePrescription Systems, and Appointment Booking. Gravitate-Health has defined personas with diverse use profiles and medication needs. Elements of the personas and associated scenarios will be used to gradually increase the complexity of the ePI scripts, while introducing step-by-step the G-lens® functionality. Among open standards, Gravitate-Health focuses on HL7 FHIR®, Web Accessibility Guidelines (WCAG2.1), and patient outcome measures like ICHOM. The G-lens® can be thought of as a set of rules that relates information in the user profile and one's patient summary to what is highlighted in ePI, or how the information from multiple ePIs is presented. HL7 FHIR® involves numerous resources and profiles. HL7 FHIR® profiles are adaptations of FHIR® resources to specific needs described in Implementation Guides (IGs). Typically, HL7 FHIR® IGs are complemented with (a) source code repositories in GitHub and (b) FHIR® servers that are populated with sample data. The Gravitate-Health IG will detail the HL7 FHIR® resources used to realize the G-lens®. The methodology adopted comprises this iterative process:

1. Identify test scenarios that mimic common real-life situations.
2. Select the FHIR® resources in-scope.
3. Identify product(s) and the SMPC/PL structured or semi-structured content
4. Convert to FHIR shorthand instances and deliver to Github workflow, to IG, to renderings of the content to HTML, XML and JSON.
5. Test at the FCAT ePI track and repeat.

The agile methodology adopted by Gravitate-Health is to progressively develop, test, and validate the resources in the IG based on FHIR shorthand (FSH), Github workflows and FHIR Release 5.0.0-snapshot1. The maturity model of HL7 FHIR® associates a level to each FHIR® resource corresponding to how extensively the resource has been tested, adopted, and used in the community. This is in line with the Gravitate Health agile methodology that entails joining the HL7 Vulcan accelerator with a dedicated ePI project, and working in a FCAT ePI dedicated track in regular sprints to incrementally advance and gradually mature the functionality of the Gravitate Health IG.

3. Results

Gravitate-Health initiated collaboration with the Vulcan Accelerator (https://confluence.hl7.org/display/VA), seeking to connect clinical research and healthcare. Gravitate-Health joined Vulcan through its coordinator early in 2021, Vulcan approved the ePI project in Aug 2021. From September 2021, Gravitate-Health committed to organizing a dedicated track in each of the three FCAT organized annually. The dedicated Vulcan ePI track develops and tries out ePI scenarios using synthetic data. In addition to the Gravitate-Health team members participating in each FCAT, all FCAT participants are invited to join and help validate the latest developments in the Gravitate-Health FHIR® IG. Parallel to each FCAT, a LinkedIn® event helps share the progress made more widely. The key ePI principles of EMA [1], advocate for "accessibility by design", so that medicinal product information is accessible to people with print impairments e.g., physical or sensory impairments, or learning difficulties. For this purpose, voice and video formats should be considered by regulators to complement

printed information and facilitate patient accessibility needs, to support use of preferred modality to perceive, understand, navigate and interact with the Internet. Thus, in terms of accessibility standards, we need to examine: (a) accessibility regulations applicable in the EU; (b) readability guidance provided by medicinal agencies applicable to patient leaflets [3]; (c) global accessibility standards; (d) accessibility of structured definitions in HL7 FHIR. The EU in the guidance already provided highlights the use of adequate font size, appropriate capitalization, uniform heading style, and line space. It also recommends short bullet lists, short sentences, and actionable guidance. A sample protocol is provided for assessing the understanding of the content by the users of the patient leaflet. Preliminary results from FCAT28 and FCAT29 are very encouraging, showing a steady increase in the participation in Gravitate-Health ePI track. Attendance in the associated LinkedIn® event also increased. Active participation in the hand-on sessions increased as well, along with the availability of tools, servers, and sample data. In fact, the use of FHIR tools like FSH and clinFHIR helped accelerate the development of sample data. **Figure 2** shows a FHIR ePI use case demonstration with CAPABLE one of the four applications that tested elements of the Gravitate-Health FHIR® IG and demonstrated elementary functionalities of the G-lens®, highlighting or suppressing ePI sections based on patient information. The ePI information is retrieved from Felleskatalogen the official product dictionary in Norway. Manually extracted knowledge linked ePI sections (*FHIR ClinicalUseIssue)* is coded in ICPC-2 with Patient demographics, allergy and conditions in *AllergyIntolerance* and *Condition* resources.

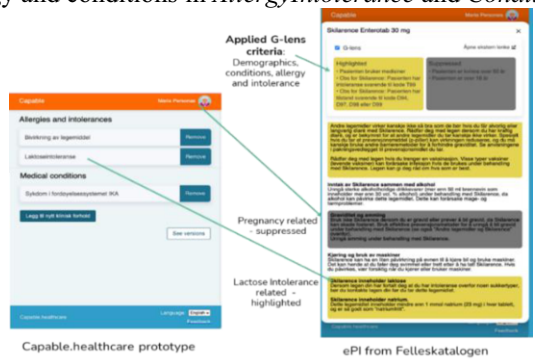


Figure 2. Example prototype from CAPABLE [5] developed in the course of FCAT29 (January 2022).

4. Discussion

The key ePI principles [1] identify ePI as a public health priority that can improve access, support adverse-event report, provide automatic update notifications and lead to more efficient regulatory processes, safety and inclusion of all. Although ePI will coexist with paper, paper and ePI will point to the newest version available in all European languages. Safeguarding these key ePI principles, Gravitate-Health collaborate with EMA, FDA, and selected national regulators to advance interoperability of resources in iterative sprint cycles and open collaboration leading to global alignment. After analyzing the landscape of standards for interoperability and accessibility, organizing two FCAT tracks, and exploring the EMA common ePI standard, there are five elements to be explored in future FCATs. First**,** knowledge representation of the ePI structured information and dictionaries connecting the elements of the user profiles and the patient summary

components is key to sustainable progress. Second, PL style and readability guidance [3] should be assessed for ability to meet the future needs for ePI and G-lens® implementations with comprehensive and real-world user scenarios with all relevant stakeholders, perhaps focusing on specific groups of pharmaceutical products. Appropriate WCAG2.1 compliant style sheets should to cater to the accessibility needs of ePI users. Third, Gravitate-Health should incorporate testing and validation guidance of EU readability and WCAG2.1, to demonstrate real-world testing of the understanding of ePIs with G-lens® use. Fourth, the federated open-source platform of Gravitate-Health could develop tools to curate and validate ePIs resources and their stylesheets, to be tested in future FCATs. Gravitate-Health should document the maturity of HL7 FHIR® ePI resources, as strategy to facilitate global adoption and regulatory alignment.

5. Conclusions and Next Steps

Interoperability driven by the HL7 FHIR® standard is at the core of the Gravitate-Health project. HL7 FHIR® is a living standard comprising of resources. The resources of interest to Gravitate-Health relate to ePI, SPOR, the International Patient Summary (IPS), and patient outcome measures. These FHIR® resources are currently at different levels of maturity and Gravitate-Health agile methodology can help accelerate their process of maturity. In addition, the open process to engage stakeholders and software developers can contribute to capacity business and community knowledge, collaborating with HL7 Vulcan accelerator for clinical research. Gravitate-Health comes into play with an agile methodology across the life cycle of ePIs that assesses understanding in the upstream and personalization in the downstream. Still there are many challenges to overcome to bring G-lens® to reality, including knowledge representation, multiple HL7 FHIR® versions, and terminology management. We expect that the agile approach adopted by Gravitate-Health will steadily move us towards advanced interoperability that drives access and understanding of medication information from trusted sources.

Acknowledgement

The project has received funding from IMI2 under Grant Agreement No:945354.

References

[1] European Medicines Agency. 2020. Electronic product information for human medicines in the EU: key principles. https://www.ema.europa.eu/en/documents/regulatory-procedural-guideline/electronic-product-information-human-medicines-european-union-key-principles_en.pdf

[2] European Medicines Agency. 2021. Action plan related to the European Commission's recommendations on product information, https://www.ema.europa.eu/en/documents/other/european-medicines-agency-action-plan-related-european-commissions-recommendations-product_en.pdf

[3] European Commission, Guideline on the readability of the labelling and packages leaflet of medicinal products, 2009 https://ec.europa.eu/health/system/files/2016/11/2009_01_12_readability_guideline_final _en_0.pdf

[4] European Medicines Agency. EU Network Data Board https://www.ema.europa.eu /en/documents/other/european-union-network-data-board-terms-reference_en.pdf (accessed 20.3.2022)

[5] Janson AL, Moen A, Fuglerud KS. Design of the CAPABLE Health Empowerment Tool: Citizens' Needs and Expectations. Stud Health Technol Inform. 2020 Jun 16;270:926-930.

Challenges of Trustable AI and Added-Value on Health
B. Séroussi et al. (Eds.)

doi:10.3233/SHTI220548

Gamification and Coaching in Remote Monitoring and Care Platforms

Antonios PARDOS[a,b], Andreas MENYCHTAS[a,b] Parisis GALLOS[a,b], Christos PANAGOPOULOS[b] and Ilias MAGLOGIANNIS[a,1]
[a] *Computational Biomedicine Research Lab, Department of Digital Systems, University of Piraeus, Greece*
[b] *Bioassist S.A., Greece*

Abstract. Nowadays, several e-health systems are equipped with advanced features for patients monitoring and care. Among these features, gamification and operations supporting the patients' adherence to therapeutic and care plans have been found to be quite useful and valuable. Among others, the introduction of intelligent patient coaching and the provisions of recommendations are very popular. The aim of this paper is to present specific gamification and coaching approaches that could be employed in the context of an existing eHealth system for remote monitoring and care for elders. The "Points, Badges and Leaderboards" gamification approach was followed. Specifically, parameters related to the application usage (daily points), the physical activity (number of daily steps), the sleep quality (sleep score) and other measurements (i.e. weight) were utilized to accommodate elders needs for motivation and engagement. Regarding the coaching, motivational messages and notification for the mobile devices were selected to deliver the relative information to the elders. A prototype health information system with a corresponding mobile application was adapted to include gamification and coaching features to motivate elders in order to achieve the maximum adherence on their monitoring and care health plans. The paper presents the design issues and summarizes the technical details.

Keywords. Gamification, PBL, coaching, elders, eHealth, mHealth, IT systems

1. Introduction

During the last decade, mobile health technologies and wearable devices have been broadly used for physical activity tracking and for the promotion of the people's well-being [1,2]. In addition, several systems have been developed lately for chronic patients and senior adults to support their care needs, to offer higher quality of life, and to promote patient independence [3-6].

Meanwhile, in the last few years, a lot of health information systems tend to support advanced features related to the prevention of diseases and the promotion of healthy lifestyles [7-10]. Among these features, gamification and operations supporting the patients' adherence to the healthcare plan have been found to be useful and valuable in patients monitoring and care [11-14]. Points - Badges - Leaderboards (PBL) gamification approach has been recently used in such prototype e-health systems revealing some promising outcomes [15].

[1] Corresponding Author, Ilias Maglogiannis, Department of Digital Systems, University of Piraeus, 80, M. Karaoli & A. Dimitriou St., 18534 Piraeus, Greece; E-mail: imaglo@unipi.gr.

Moreover, the concept of intelligent (autonomous) patient coaching in modern healthcare information systems is very popular and acceptable, since the individual provision of information, advice and incentives can offer a richer, safer, and more effective user experience. The intelligent patient coaching can be characterized as a valuable tool in the remote care systems as it can provide multiple advantages for the patient, such as defining individual care plans or suggestions for each user while adapting them according to the patient's performance or preferences. Based on the international literature, coaching can be delivered using motivational messages [16] which can improve the user's confidence, while could develop a rapport between the system and the user [17,18]. In addition, the push notifications technology in mobile devices have also been found to be suitable for the coaching support [19]. The content of both motivational messages and notifications can be either informative on general topics (for example, the user can be updated or informed regarding healthcare issues while he/she takes a break from his/her daily activities) or could be adapted to the individual program specified for a specific user based on his/her personal performance to receive personalized information.

The aim of this paper is to present the gamification and coaching approaches that could be used in the context of an existing eHealth system for remote monitoring and care for elders and chronic patients. The introduction of the gamification and coaching features on eHealth systems may lead to the increase of patients' motivation to adapt their behaviour for a healthier lifestyle and improved quality of life.

2. Methods

As already mentioned, in order to implement a gamification feature on an eHealth platform, the PBL gamification approach can be followed. Specifically, parameters relative to the application usage (daily points), the physical activity (number of daily steps), the sleep quality (sleep score) and other several measurements (ex. weight) can be utilized to accommodate elders needs for motivation and engagement.

In the initial design of the envisaged system, one of the requirements is the ability to record the aforementioned factors for each user. The above factors' data analysis [15] produce specific points based on the user's performance. The collection of the points leads to the user award with badges. The number of the different badges per factor and the required points for each badge can be defined by the system administrator in collaboration with the healthcare professionals who can assess the importance of each badge. Another requirement of the system includes the information of the user for his/her rank among the other users of the system. The rank is calculated based on the collected points on the above factors. Specifically, functions such as *min*, *max, sum, count* can be applied on raw collected data for points calculation. In addition, the collected data may be also combined with the measurements' time to produce more complex badge types.

Regarding the coaching requirements included, motivational messages feature was selected as an informative approach to support and motivate the elders in combination with the above-mentioned gamification approach. Finally, the usage of notification feature for the mobile app users was also be considered as an available option to enhance the elders coaching.

3. Results and Discussion

The gamification and coaching features were developed with a modular based architecture and were integrated with a homecare platform for validation. The platform interface was redesigned, to be friendlier and understandable to elders. Figure 1 presents the "My Badges" screen where the user can be informed about the earned badges and the points on each task/mission/factor. The points calculation is held using cloud computing and relative microservices that are followed by the platform. The current prototype includes badges for the application usage by measuring the frequency of use, the daily physical activity by analysing the number of steps, the sleep quality by analysing the sleep score and the user's weight. All the above data can be collected by smart devices such as smartwatches or fit trackers and smart scales that are compatible with the platform and support direct or indirect data exchange.

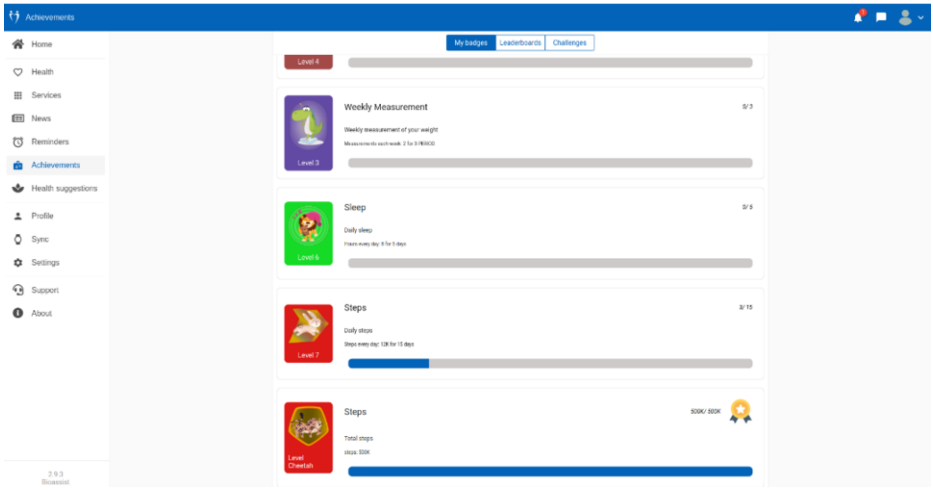


Figure 1. The "My Badges" screen

In order to increase the user's motivation and to promote social aspects of the system the gamification leaderboard and comparison graphs were developed. Figure 2 shows the user's performance comparison graphs. Each user can locate his/her place on the leaderboard and can also find the other users' performance for the same task/mission.

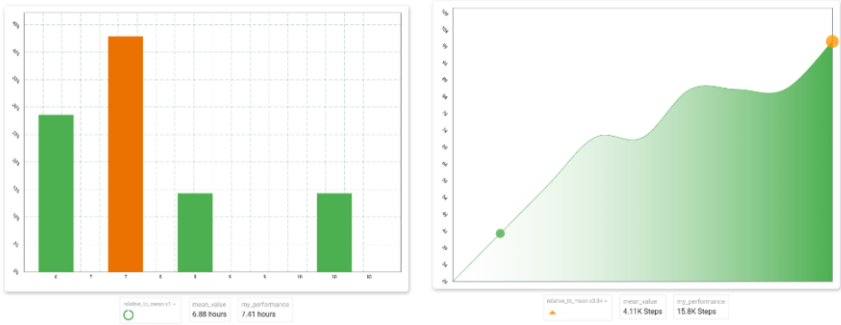


Figure 2. Patients' Performance Comparison Graphs

Using the figure 2 graphs, users can be informed regarding their performance for several factors. Their performance comparison with other similar users can lead them to adjust their behaviours in order to "win the game" and fulfil their targets.

Personalised motivation messages feature was also developed and used in combination with the platform's notification function. Based on the user's rank on the leaderboard, his/her condition, personal characteristics and activity, the elders can receive through the mobile application personalised motivation messages to continue the suggested by their care plan activities. The application of gamification and serious games for personalized health has been previously discussed on other relative studies [20].

Several health information systems have been used lately having similar features based on specific gamification or coaching approaches with very promising results. Gamification and Coaching functionalities seem to be incorporated with Electronic Health Records (EHR) to produce more accurate calculation on target setting and better ranking of the patients at leaderboards [21,22]. On the other hand, the usage of large EHR and the formulation of medical data silos to support the above functionalities arises issues related to the health data remote control and access by different eHealth platforms. Other relative studies present the usage of gamification for smoking cessation [23] or older adults support [24] and the strong relation with the mHealth apps [12, 25]. The relative advantage of the proposed system is the combination of the gamification features with the motivation messaging functionalities and the ranking of the users in dynamic leader boards.

4. Conclusions

A prototype health information system with a relative mobile application were adapted to include gamification and coaching features to motivate elders in order to achieve the maximum adherence on their monitoring and care health plans. The proposed solution is fully flexible and follows modular design as can be applied and integrated in any similar system. Limitations of the current work include the modules development based on the requirements for specific factors. Also, the presented prototype's integration with an real-world care plan and its role in the shared decision making process are not examined yet. Further examination and introduction of other related factors as well as extensions with additional user competitions are included on the project's future work. In addition, possible implementation of the proposed coaching feature in Electronic Health Record systems and Remote Care systems where healthcare professionals access the patient's holistic health records and they can set new targets, are also some of the future ideas for expandability and exploitation.

Acknowledgements

This research has been co-financed by the European Union and Greek national funds through the Operational Program Competitiveness, Entrepreneurship and Innovation, under the call RESEARCH – CREATE – INNOVATE (project code: MediLudus Personalised home care based on game and gamify elements T 2EKΔK-03049).

References

[1] Romeo A, Edney S, Plotnikoff R, Curtis R, Ryan J, Sanders I, et al. Can Smartphone Apps Increase Physical Activity? Systematic Review and Meta-Analysis. J Med Internet Res. 2019 Mar 19;21(3):e12053. doi: 10.2196/12053. PMID: 30888321; PMCID: PMC6444212.

[2] Hutchesson MJ, Gough C, Müller AM, Short CE, Whatnall MC, et al. eHealth interventions targeting nutrition, physical activity, sedentary behavior, or obesity in adults: A scoping review of systematic reviews. Obes Rev. 2021 Oct;22(10):e13295.

[3] Demiris G, Thompson H. Smart homes and ambient assisted living applications: from data to knowledge-empowering or overwhelming older adults? Contribution of the IMIA Smart Homes and Ambiant Assisted Living Working Group. Yearb Med Inform. 2011;6:51-7. PMID: 21938325.

[4] Panagopoulos C, Menychtas A, Tsanakas P, Maglogiannis I. Increasing Usability of Homecare Applications for Older Adults: A Case Study. Designs. 2019; 3(2):23

[5] Vimarlund V, Wass S. Big data, smart homes and ambient assisted living. Yearb Med Inform. 2014 Aug 15;9(1):143-9. doi: 10.15265/IY-2014-0011. PMID: 25123734; PMCID: PMC4287073.

[6] Maeder A. Digital Technology Trends Supporting Assisted Independent Living of Ageing Population. Stud Health Technol Inform. 2020 Jun 16;270:577-581. doi: 10.3233/SHTI200226. PMID: 32570449.

[7] DiPietro J, et al. Computer- and Robot-Assisted Therapies to Aid Social and Intellectual Functioning of Children with Autism Spectrum Disorder. Medicina (Kaunas). 2019 Aug 5;55(8):440.

[8] Panagopoulos C, et al. Utilizing a Homecare Platform for Remote Monitoring of Patients with Idiopathic Pulmonary Fibrosis. Advances in Experimental Medicine and Biology. 2017;989.

[9] Menychtas A, Doukas C, Tsanakas P, Maglogiannis I. A Versatile Architecture for Building IoT Quantified-Self Applications. 2017 IEEE 30th International Symposium on Computer-Based Medical Systems (CBMS). 2017;500-505

[10] Kallipolitis A, et al. Affective analysis of patients in homecare video-assisted telemedicine using computational intelligence. Neural Comput & Applic. 2020;32:17125–17136.

[11] Edwards EA, Lumsden J, Rivas C, Steed L, Edwards LA, Thiyagarajan A, et al. Gamification for health promotion: systematic review of behaviour change techniques in smartphone apps. BMJ Open. 2016 Oct 4;6(10):e012447.

[12] Miller AS, Cafazzo JA, Seto E. A game plan: Gamification design principles in mHealth applications for chronic disease management. Health Informatics J. 2016 Jun;22(2):184-93.

[13] Sardi L, Idri A, Fernández-Alemán JL. A systematic review of gamification in e-Health. J Biomed Inform. 2017 Jul;71:31-48. doi: 10.1016/j.jbi.2017.05.011. Epub 2017 May 20. PMID: 28536062.

[14] Cheng VWS, Davenport T, Johnson D, Vella K, Hickie IB. Gamification in Apps and Technologies for Improving Mental Health and Well-Being: Systematic Review. JMIR Ment Health. 2019;6(6):e13717.

[15] Pardos A, Menychtas A, Maglogiannis I. Introducing Gamification in eHealth Platforms for Promoting Wellbeing. Stud Health Technol Inform. 2022 Jan 14;289:337-340

[16] Heffernan KJ, Chang S, Maclean ST, Callegari ET, Garland SM, Reavley NJ, et al. Guidelines and Recommendations for Developing Interactive eHealth Apps for Complex Messaging in Health Promotion. JMIR Mhealth Uhealth. 2016 Feb 9;4(1):e14.

[17] Ding L, Kim C, Orey M. (2017). Studies of student engagement in gamified online discussions. Computers & Education, 115, 126-142.

[18] Lee TY, Dugan C, Geyer W, Ratchford T, Rasmussen J, Shami NS, et al. (2013, June). Experiments on motivational feedback for crowdsourced workers. In Seventh International AAAI Conference on Weblogs and Social Media.

[19] Kazhamiakin R, Marconi A, Martinelli A, Pistore M, Valetto G. (2016, September). A gamification framework for the long-term engagement of smart citizens. In 2016 IEEE International Smart Cities Conference (ISC2) (pp. 1-7). IEEE.

[20] McCallum S. Gamification and serious games for personalized health. Stud Health Technol Inform. 2012;177:85-96.

[21] Gibbs D, Hewitt B, McLeod A. Gamification of Electronic Health Records: A Systematic Literature Review. Educational Perspectives in Health Informatics and Information Management. 2016.

[22] Andrikopoulou E, Scott P, Herrera H, Good A. What are the important design features of personal health records to improve medication adherence for patients with long-term conditions? A systematic literature review. BMJ Open. 2019 Sep 26;9(9):e028628.

[23] Pløhn T, Aalberg T. (2015, October). Using gamification to motivate smoking cessation. In European Conference on Games Based Learning (p. 431). Academic Conferences International Limited.

[24] Tabak M, et al. A Game-Based, Physical Activity Coaching Application for Older Adults: Design Approach and User Experience in Daily Life. Games Health J. 2020 Jun;9(3):215-226.

[25] Schmidt-Kraepelin M, Toussaint PA, Thiebes S, Hamari J, Sunyaev A. Archetypes of Gamification: Analysis of mHealth Apps. JMIR Mhealth Uhealth. 2020 Oct 19;8(10):e19280.

Challenges of Trustable AI and Added-Value on Health
B. Séroussi et al. (Eds.)

doi:10.3233/SHTI220549

Implementing SNOMED CT in Open Software Solutions to Enhance the Findability of COVID-19 Questionnaires

Carina Nina VORISEK[a,1], Eduardo Alonso ESSENWANGER [a], Sophie Anne Ines KLOPFENSTEIN[a,b], Julian SASS[a], Jörg HENKE[c], Carsten Oliver SCHMIDT[c], Sylvia THUN[a], on behalf of the NFDI4Health Task Force COVID-19
[a]*Berlin Institute of Health at Charité – Universitätsmedizin Berlin*
[b]*Institute of Medical Informatics, Charité – Universitätsmedizin Berlin, Berlin*
[c]*University Medicine of Greifswald*

Abstract. SNOMED CT fosters interoperability in healthcare and research. This use case implemented SNOMED CT for browsing COVID-19 questionnaires in the open-software solutions OPAL/MICA. We implemented a test server requiring files in a given YAML format for implementation of taxonomies with only two levels of hierarchy. Within this format, neither the implementation of SNOMED CT hierarchies and post-coordination nor the use of release files were possible. To solve this, Python scripts were written to integrate the required SNOMED CT concepts (Fully Specified Name, FSN and SNOMED CT Identifier, SCTID) into the YAML format (YAML Mode). Mappings of SNOMED CT to data items of the questionnaires had to be provided as Excel files for implementation into Opal/MICA and further Python scripts were established within the Excel Mode. Finally, a total of eight questionnaires containing 1.178 data items were successfully mapped to SNOMED CT and implemented in OPAL/MICA. This use case showed that implementing SNOMED CT for browsing COVID-19 questionnaires is feasible despite software solutions not supporting SNOMED CT. However, limitations of not being able to implement SNOMED CT release files and its provided hierarchy and post-coordination still have to be overcome.

Keywords. SNOMED CT, Semantic Interoperability, Standardization, COVID-19

1. Introduction

The NFDI4Health Task Force COVID-19, an interdisciplinary German network project within the National Research Data Infrastructure for Personal Health Data (NFDI4Health) initiative is aiming for a better overview on public health, epidemiological and clinical studies targeting the current pandemic [1]. Therefore, three interlinked platforms were developed containing: 1) a Central Search Hub for study level information, 2) a COVID-19 study hub SEEK platform, linking COVID-19 studies with their metadata, documents (assets) and other information, 3) and Opal and MICA, containing items from selected COVID-19 survey instruments [2]. OPAL/MICA are open software solutions built for managing and harmonizing epidemiological data [3].

Using a common terminology can improve interoperability and secondary use of data within data infrastructures. Currently, SNOMED CT is the most comprehensive clinical healthcare terminology worldwide providing more than 350.000 concepts. Its organization SNOMED International comprises 41 member countries [4]. The three

components “concepts”, “descriptions” and “relationships” enable SNOMED CT’s features such as compositional grammar, expression constraint queries and post-coordination offering possibilities to combine SNOMED CT concepts [5]. In MICA, a simple semantic structure to classify study items is available, the Maelstrom taxonomy, which is composed of 18 domains and 135 subdomains [6].

To provide a faceted search of COVID-19 questionnaires, we aimed to implement SNOMED CT in addition to the already used Maelstrom taxonomy into MICA/OPAL, to further improve the findability of COVID-19 questionnaires. This study evaluates whether SNOMED CT can be implemented into open software solutions, which were initially not focused on the support of SNOMED CT.

2. Methods

As a first step, we mapped SNOMED CT concepts to data items of COVID-19 questionnaires already stored in OPAL/MICA as described previously [7]. Next, a test server based on OPAL/MICA was implemented for our use case. As OPAL/MICA requires files in a given YAML format for implementation of taxonomies and allows only two levels of hierarchy, using SNOMED CT release files for import was not possible. Therefore, a Python script was written to integrate the required SNOMED CT concepts (Fully Specified Name, FSN and SNOMED CT Identifier, SCTID) into the YAML format (YAML Mode). In addition, OPAL/MICA require specific Excel file formats for implementing questionnaires and their mappings to standards which is targeted by the Excel mode of the Python Script. Mapping files, Excel formats and the Python script are provided on GitHub [8].

3. Results

3.1. Mapping to SNOMED CT

A total of eight questionnaires containing 1.178 data items were mapped to SNOMED CT and then stored in OPAL/MICA. OPAL/MICA require specific Excel file formats allowing only one SNOMED CT concept mapped to one data item. The mapping limitations are presented in detail in **table 1**.

Table 1. Limitations of Mapping SNOMED CT concepts to Data Items of the Questionnaires

Data Item	Does the patient suffer from a cardiovascular disease? History of heart revascularization
Current mapping in OPAL/MICA	Disorder of cardiovascular system (disorder)
Possible mapping using post-coordination	|Past history of procedure (situation)|:|Associated procedure (attribute)|=|Heart revascularization (procedure)|

3.2. Python Script - Excel mode:

The Excel mode (**Figure 1)** read the */Files* folder and looped over all its Excel files. Every file contained "FSN_X" and "SCTID_X" columns (X = non-negative value). The script filtered out the semantic tags listed on the "FSN_X" column and created a "SNOMED::X" column with "X" depending on the semantic tag for all semantic tags. The prefix SNOMED is defined in the taxonomy definition given to OPAL/MICA that is the precondition before any mappings can be uploaded completely. The values of the "SCTID_X" column was mapped to the corresponding "SNOMED::X" column. Finally, the script created Excel files in the required format in the */Output* folder.

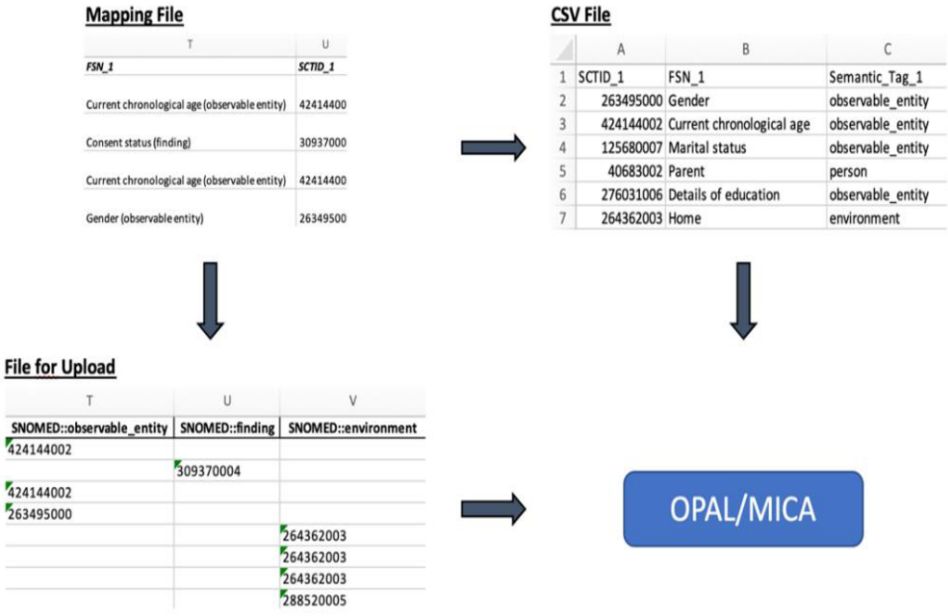


Figure 1. Shows the Workflow of the Excel Mode in the Python Script.

3.3. Python Script - YAML mode:

This mode created the YAML file (**Figure 2**) in a given format required by OPAL/MICA. Excel files located in the */Files* folder were read and merged into one CSV file with the following headers "SCTID_X"; "FSN_X"; "Semantic_Tag_X". The YAML file was created together with the csv2yaml.py file and finally saved in the */Output* folder.

Figure 2. Shows the Workflow of the YAML Mode in the Python Script.

3.4. Display of Results in OPAL/MICA

User experience was improved by categorizing the SNOMED CT concepts by their semantic tag utilizing the two level hierarchy option in OPAL/MICA. SCTIDs were provided in the background, therefore browsing was possible using the FSN as well as the SCTID. Implementation results in OPAL/MICA are provided in **figure 3**.

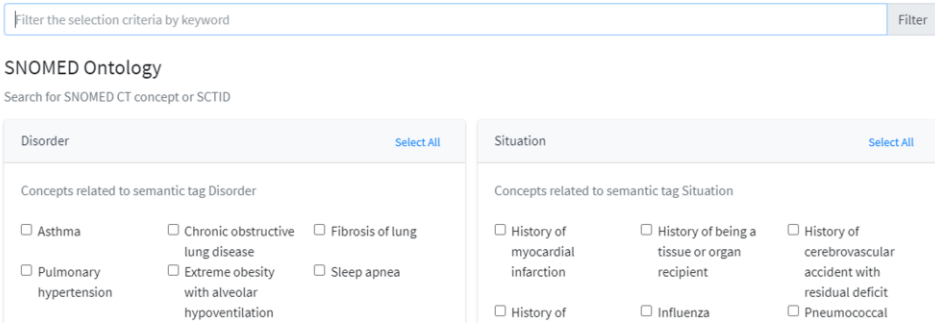


Figure 3. SNOMED CT displayed in OPAL/MICA.

4. Discussion and Conclusion

To our knowledge, faceted search using SNOMED CT in open software solutions has not been implemented nor tested. This use case showed that implementing SNOMED CT for browsing questionnaires related to COVID-19 studies is feasible despite software solutions initially not designed to support SNOMED CT. However, we faced several limitations using OPAL/MICA: We needed to adhere to specific formatting requirements allowing only one SNOMED CT concept per data item consequently inhibiting

SNOMED CT's grammar and post-coordination. Due to the required YAML format when importing new taxonomies into OPAL/MICA, the use of SNOMED CT's release files was not possible. Converting between these different file formats manually fostered slow working processes, enormous workload and most likely human error.

Therefore, Python Scripts were written to transfer the Excel files established during the mapping process into Excel files implementing the annotated questionnaires (Excel Mode) and SNOMED CT as taxonomy into OPAL/MICA (YAML Mode). With the help of our Python Scripts, we created an automated workflow from Mapping to Implementation. To further integrate SNOMED CT features such as querying, grammar and post-coordination into OPAL/MICA, the software applications need to be adapted and terminology servers need to be established.

This study demonstrates the implementation of SNOMED CT in open software solutions often used in academic institutions, smaller healthcare facilities as well as low- and middle-income countries with limited financial opportunities. Therefore, our novel approach using terminologies in open-source environment can enhance the success of their implementation. In addition, we aim to extend our use-case beyond COVID-19 studies and hence elevate the overall findability of study items from clinical, epidemiological and public health studies.

5. Acknowledgements

This work was done as part of the NFDI4Health Task Force COVID-19 (www.nfdi4health.de). We gratefully acknowledge the financial support of the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – Project Number 451265285, PI 345/17-1/SCHM 2744/9-1.

References

[1] Task Force COVID-19 - NFDI4Health. https://www.nfdi4health.de/de/task-force-covid-19/ (accessed August 11, 2021).
[2] Schmidt CO, Darms J, Shutsko A, et al. Facilitating Study and Item Level Browsing for Clinical and Epidemiological COVID-19 Studies, Stud. Health Technol. Inform. 281 (2021) 794–798.
[3] Doiron D, Marcon Y, Fortier I, et al. Software Application Profile: Opal and Mica: open-source software solutions for epidemiological data management, harmonization and dissemination., Int. J. Epidemiol. 46 (2017) 1372–1378.
[4] SNOMED - Members. https://www.snomed.org/our-customers/members (accessed January 18, 2022).
[5] SNOMED CT Starter Guide - SNOMED CT Starter Guide - SNOMED Confluence. https://confluence.ihtsdotools.org/display/DOCSTART/SNOMED+CT+Starter+Guide?preview=/28742871/47677485/doc_StarterGuide_Current-en-US_INT_20170728.pdf (accessed January 18, 2022).
[6] Bergeron J, Doiron D, Marcon Y, et al. Fortier, Fostering population-based cohort data discovery: The Maelstrom Research cataloguing toolkit, PLoS One. 13 (2018) e0200926.
[7] Vorisek CN, Klopfenstein SI, Sass J, et al. Evaluating Suitability of SNOMED CT in Structured Searches for COVID-19 Studies, Stud. Health Technol. Inform. 281 (2021) 88–92.
[8] BIH-CEI/SNOMEDCT_OPAL. https://github.com/BIH-CEI/SNOMEDCT_OPAL (accessed January 18, 2022).

Challenges of Trustable AI and Added-Value on Health
B. Séroussi et al. (Eds.)

doi:10.3233/SHTI220550

Dynamic Prediction of Non-Neutral SARS-Cov-2 Variants Using Incremental Machine Learning

Giovanna NICORA[a,b], Simone MARINI[c], Marco SALEMI[d] and Riccardo BELLAZZI[a]
[a]*Dept. of Electrical, Computer and Biomedical Engineering, University of Pavia, Italy*
[b]*enGenome S.r.l, Via Ferrata 5, Pavia, Italy*
[c]*Dept. of Epidemiology, University of Florida, Gainesville (FL)*
[d]*Dept. of Pathology, University of Florida, Gainesville (FL)*

Abstract. In this work we show that Incremental Machine Learning can be used to predict the classification of emerging SARS-CoV-2 lineages, dynamically distinguishing between neutral variants and non-neutral ones, i.e. variants of interest or variants of concerns. Starting from the Spike protein primary sequences collected in the GISAID db, we have derived a set of k-mers features, i.e., aminoacid subsequences with fixed length *k*. We have then implemented a Logistic Regression Incremental Learner that was monthly tested on the variants collected since February 2020 until October 2021. The average value of balanced accuracy of the classifier is 0.72 ± 0.2, which increased to 0.78 ± 0.16 in the last 12 months. The alpha, beta, gamma, eta, kappa and delta variants were recognized as non-neutral variants with mean recall ~90%. In summary, incremental learning proved to be a useful instrument for pandemic surveillance, given its capability to update the model on new data over time

Keywords. SARS-Cov-2 variant prediction, Incremental Machine Learning

1. Introduction

The current pandemic caused by the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) has seen the progressive emergence of several relevant variants. The Centers for Disease Control and Prevention (CDC) has classified existing SARS-CoV-2 lineages into Neutral variants, variants of interest (VOI) and variants of concern (VOC) [1]. VOI are variants with specific genetic markers that have been associated with receptor binding change, reduced neutralization by antibodies and treatments efficacy, potential diagnostic impact, predicted increase in transmissibility or disease severity. VOCs are VOI variants for which there is evidence of an increase in transmissibility, more severe disease, significant reduction in neutralization by antibodies generated during previous infection or vaccination, reduced effectiveness of treatments or vaccines, or diagnostic test failures. Variants are classified after their characteristics have been assessed from a public health point of view, such as, enhanced transmissibility. For this reason, the counter measures are always implemented after a variant has been classified, i.e., the virus has always the upper hand in the arms race against the variants. Consequently, recognising a VOI or VOC as early as possible is utterly important to curb its damage, and ultimately save lives. Machine learning (ML) has been widely applied

to COVID-19 data. Syeda et al. [2] showed most of the studies focused on outbreak prediction, drug discovery or repurposing, COVID-19 detection from patient data, disease progression and outcomes predictions. Only few studies were devoted to variant-related predictions, for example in isolating critical amino acid positions (or patterns) in the spike protein [3], or in forecasting novel variant potential waves [4]. Of note, the Pango Lineages framework has a specific ML module (PangoLearn) that classify unknown viral genomes into already known lineages. Different from the aforementioned approaches, here we propose to use ML to predict VOI and VOC, including the ones that yet have to emerge. That is, we developed an algorithm predicting the classification of each variant as it emerges, ideally before it spreads enough to manifest its related phenotypes–ahead of its official classification. In order to define a strategy that can be used in a realistic context, we propose a method based on incremental learning [5], able to dynamically modify the prediction model during data collection and to adapt to the pandemic evolution.

2. Methods

Data set. Our data set consists of Spike protein primary sequences from GISAID collected on October, 2021. We decided to focus on spike protein sequences because VOC and VOI lineage classifications are based on mutations in Spike proteins; moreover, by only focusing on the Spike 1350 amino acids, we limit the feature space (as opposed to considering all the SARS-CoV-2 proteins). After removing duplicated sequences, and filtering the Spike proteins based on both the frequency of uncharacterized amino acids (AAs), set to a maximum of 1%, and length, set to a minimum of 1000 AAs, we obtained a final set of 130,772 proteins. Of these, 42156 were Neutral (32%), 3062 were VOIs (2%) and 85552 were VOCs (66%). We then grouped the VOI and VOC variants in the same "Non-Neutral" class. Figure 1 shows the number of Neutral and Non-Neutral sequences collected from February 2020 to October 2021. Not surprisingly, at the beginning of the pandemic, most of the sequences deposited in GISAID were Neutral, while from April, 2021, the number of Non-neutral variants highly increased.

Feature representation. We translated proteins sequences into a fixed-length set of numeric features through k-mers, so that each protein, independently from its length, will have a numeric representation. K-mers are a classical method to represent biological nucleic or amino acid sequences, widely used in bioinformatics [6]. Briefly, k-mers are substrings of user-defined length k contained in a sequence. For example, given k=2, we find in the sequence GATTACA the k-mers 'GA', 'AT', 'TT', 'TA', and 'CA'. Each k-mer has a boolean value indicating its presence/absence. Since we wanted to represent variations of one to few amino acids, we considered small ks, i.e. k=3 and k=5. We removed k-mers containing the "X" character, indicating a missing value.

Incremental Learner Representation. To simulate the implementation of a pandemic surveillance classifier that is updated over time, we make use of incremental learning. Incremental learning algorithms are able to incorporate new knowledge without a complete re-training of model parameters [5]. We simulated a scenario with a one-month pace, as follows: in March 2020, the very first set of spike proteins deposited in GISAID up to February 2020 are used to train the classifier. Only 73 Neutral sequences were on GISAID at the time (Figure 1). Therefore, our classifier was first be trained on a small number of negative (i.e. Neutral) variants. On April 2020, the new sequences deposited in March are available for classification. Our classifier predicts the class of the new

variants based on the training from the previous month, and the performances are recorded. Then, supposing that the true class (Neutral vs Non-Neutral) is available, the incremental classifier is updated by partially re-training on the new set of data (March 2020). On May 2020, the same procedure is applied on sequences collected along April 2020, and so on, up to October 2021 (19 learning steps). The incremental learner has been implemented using the SGDC Classifier available in sklearn (https://scikit-learn.org/stable/. We implemented two logistic regression incremental learners with Lasso penalty using k = 3 and k = 5. Note that for the first training (March 2020), we remove all k-mers that are present in all the sequences (frequency 1), or totally absent (frequency 0).

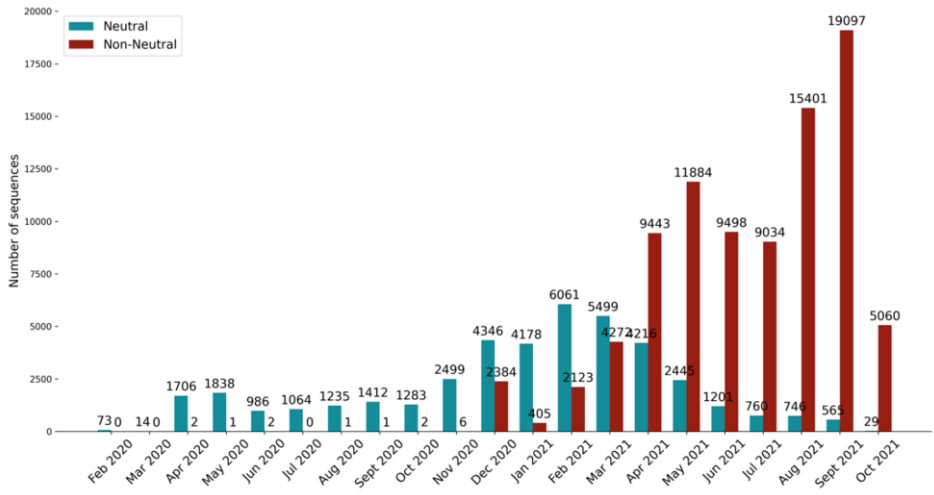


Figure 1. Number of sequences with a percentage of missing values (X) below 1%, submitted to GISAID each month from February, 2020 to October, 2021.

3. Results

For the sake of brevity, here we report the results of the model trained with kmer of length 3, since the two models showed comparable results (data not shown). After filtering, around 738 kmers are kept as features. Figure 2 shows that recall (i.e. percentage of Non-Neutral variants correctly predicted) and specificity (i.e. percentage of Neutral variants correctly predicted) changed over time as the learner was updated and tested on new chunks. It is interesting to note that the results are strongly influenced by two non stationary factors: i) the increase in the total number of variants; ii) the corresponding increase of non-neutral variants. The performance of the methods is in general satisfactory, even if the sudden changes in the proportions between classes caused a decrease in specificity (thus increasing false positives) in the last part of the considered time window. Looking the balanced accuracy, the average value of the total reported period is 0.72 ± 0.2, while in the last 12 months increased to 0.78 ± 0.16, thus showing relatively overall good results of the forecasting algorithm. We also stratified results according to the WHO labeling applied to VOC and VOI variants (Table 1). The

alpha, beta, gamma, eta, kappa and delta variants were usually well recognized as Non-neutral variants (mean recall around 90%). The learner was less sensitive to the mu and zeta variants, probably because they were less represented in the dataset.

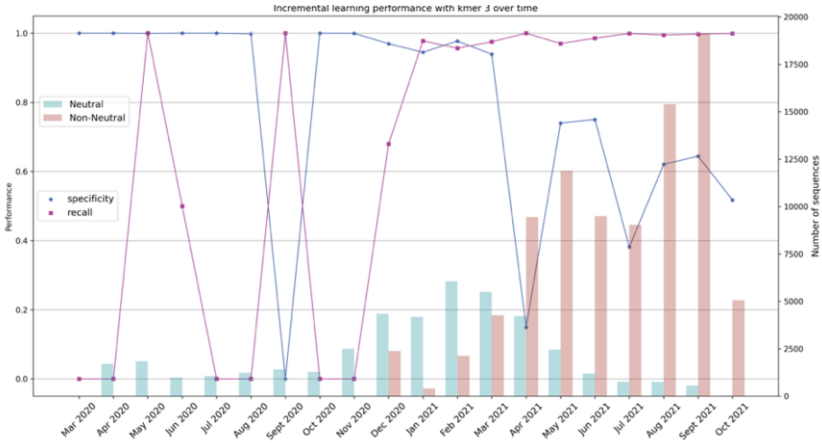


Figure 2. Sensitivities and specificities of the incremental learner trained with k=3, calculated at each month.

Table 1. Recall stratified by variant type as reported by the WHO (https://www.who.int/en/activities/tracking-SARS-CoV-2-variants/). The total recall is calculated on the entire set of VOC variants from March 2020 to October 2021. The Mean Recall is the average value of the recall values calculated at each time step (month), while the Std recall is the standard deviation of the recall values at each time step.

	Alpha	**Beta**	**Gamma**	**Epsilon**	**Eta**	**Iota**	**Kappa**	**Mu**	**Zeta**	**Delta**
Total Number	29795	1363	3385	833	509	1436	868	135	176	48343
Total Recall	99.34	95.52	98.37	95.07	94.49	89.76	99.53	56.29	64.2	98.2
Mean Recall	92.55	89.66	89.6	72.71	89.7	73.2	90.7	68.93	47.21	90.61
Std Recall	0.25	0.28	0.3	0.42	0.17	0.3	0.28	0.34	0.46	0.14

4. Discussion

During a pandemic caused by viruses such as the SARS-CoV-2, detecting new variants and understanding their effects as soon as possible is of paramount importance. We used the information encoded in spike protein sequences to incrementally train a ML classifier in predicting SARS-Cov-2 variants as Neutral or Non-Neutral. We show that our framework can be used in a real case scenario to select the most concerning variants (predicted as Non-Neutral) for further laboratory testing to assess their potential harm even before they spread. In this context, an important information is *when* a VOC is detected: even in the early phase of the pandemic, when very few VOCs are reported, if at least one of them is classified as VOC by the classifier, in vitro tests would confirm the deleteriousness of this variant, allowing for early detecting of VOC. Several aspects need to be considered when developing ML tools applied to protein sequences. First of

all, we need to find a proper numerical representation of the AA sequences. We chose to represent each protein with k-mers. Another possibility would be to use deep learning embedding. Secondly, the problem is highly imbalanced, and the number of sequences exponentially increases over time (Figure 1). Also, the number of Neutral variants decreases in time, and we foresee a specificity decrease, suggesting that dataset shift is occurring, i.e. the population of Neutral variants is changing. Therefore the model is less reliable in the classification of Neutral class. In a real setting, this information may be used to decide a re-training of the algorithm, as part of a monitoring processing. At each step (i.e., month), after testing the classifier on newly collected data, we assume that the true class of the sequences is available, and we partially refit the model on these data using the true class. Other possible approaches include, for example, to use the predicted class for training, in a semi-supervised fashion, and to completely re-train the classifier at each step. This may be feasible especially at the beginning, when few sequences are collected, but might cost a considerable computational effort later on. Instead, incremental learning allows us to update the model with a less computationally expensive approach and by controlling the weight of the most recent data. Yet, the need for manual download from GISAID to obtain the proteins and the exponentially increasing number of proteins available may be a bottleneck. Future works will analyse multi-class classifier able to distinguish between VOC, VOI and Neutral. Moreover, the new VOC labeled as Omicron will be included in the analysis. Thanks to the k-mer representation, our proposed pipeline can be adapted to other types of viruses as well.

5. Conclusion

We developed a pipeline classify new SARS-CoV-2 variants as Neutral or Non-Neutral as they emerge. Incremental learning proved to be a useful instrument for pandemic surveillance, given its capability to update the model on new data over time. Additionally, it is computationally less demanding in comparison with a complete re training of the algorithm, making it better suited to “big-data” bioinformatics applications.

References

[1] CDC. Coronavirus Disease 2019 (COVID-19); 2020. Available from: https://www.cdc.gov/coronavirus/2019-ncov/variants/variant-classifications.html.

[2] Syeda HB, Syed M, Sexton KW, Syed S, Begum S, Syed F, et al. Role of Machine Learning Techniques to Tackle the COVID-19 Crisis: Systematic Review. JMIR medical informatics. 2021 Jan;9(1):e23811.

[3]. Nagpal S, Pal R, Ashima, Tyagi A, Tripathi S, Nagori A, et al. Genomic Surveillance of COVID-19 Variants with Language Models and Machine Learning; 2021. Company: Cold Spring Harbor Laboratory Distributor: Cold Spring Harbor Laboratory Label: Cold Spring Harbor Laboratory Section: New Result Type: article. Available from: https://www.biorxiv.org/content/10.1101/2021.05.25. 445601v2.

[4] Rashed EA, Hirata A. Infectivity Upsurge by COVID-19 Viral Variants in Japan: Evidence from Deep Learning Modeling. International Journal of Environmental Research and Public Health. 2021 Jan;18(15):7799. Number: 15 Publisher: Multidisciplinary Digital Publishing Institute. Available from: https://www.mdpi.com/1660-4601/18/15/7799.

[5] Bouchachia A, Gabrys B, Sahel Z. Overview of Some Incremental Learning Algorithms. In: 2007 IEEE International Fuzzy Systems Conference; 2007. p. 1-6.

[6] Marchet C, Boucher C, Puglisi SJ, Medvedev P, Salson M, Chikhi R. Data structures based on k-mers for querying large collections of sequencing data sets. Genome Research. 2021 Jan;31(1):1-12. Available from: https://genome.cshlp.org/content/31/1/1.

Challenges of Trustable AI and Added-Value on Health
B. Séroussi et al. (Eds.)

doi:10.3233/SHTI220551

Atlantes: Automated Health Related & COVID-19 Data Management for Use in Predictive Models

George VANGELATOS[a,1], Haralampos KARANIKAS[a] and Sotiris TASOULIS [a]

[a] *Department of Computer Science and Biomedical Informatics, University of Thessaly, Lamia, Greece*

Abstract. The scientific community, having turned its interest, almost entirely, to the treatment and understanding of COVID-19, is constantly striving to collect and use data from the countless available sources. That data, however, is scattered, not designed to be combined, collected in different time periods and their volume is constantly increasing. In this paper, we present an automated methodology that collects, refines, groups and combines data for a large number of countries. Most of these data resources are directly related to COVID-19 but we also choose to include other types of variables for each country, which may be of particular interest for researchers working in understanding the COVID-19 pandemic. The presented methodology unifies critical information regarding the pandemic. It is implemented in Python, provided as a simple script that extracts data, in the form of a daily time series, in a short period of time, directly available to be incorporated for analysis.

Keywords. COVID-19, open-source, health data collection, data management

1. Introduction

In late 2019, an infection of unknown origin was detected with manifestations of respiratory diseases as it began to spread rapidly in Hubei Province, China, especially in its largest city, Wuhan. Shortly afterwards, the World Health Organization designated coronavirus 2019 (COVID-19) and by March, the outbreak was a "pandemic" [1].

Although there has been significant forecasting and research, there is little data available to actively monitor current data trends and their statistical and practical significance. In addition, in most cases this data is available from different sources, with different formats and even different units of measurement for different time periods. For this reason, in this paper, we introduce a unified methodology in the form of an open-source Python script that is able to aid in the collection of all information relevant, and / or possibly relevant, to COVID-19 and their integration into a coherent data set with universal names, units of measurement and pre-specified time frames. This way, we expect to enforce data analytics and machine learning research towards deploying appropriate prediction and forecasting tools for COVID-19 outbreaks [9]

[1] Corresponding Author, George Vangelatos, Department of Computer Science and Biomedical Informatics, University of Thessaly, Papasiopoulou 2-4, Postal Code: 35131, Lamia, Greece; E-mail: gvangelatos@uth.gr

Our methodology supports the PrescITs' project Knowledge base (KB). The PrescIT project[2] aims to develop a Clinical Decision Support System (CDSS) platform to support safe e-Prescription, using per-disease rule-based algorithms to prescribe treatment [2] and knowledge extraction from reliable sources, such as large drug databases, scientific literature and the pharmaceutical market. One future research direction that the KB could support, is the investigation of relations among the COVID-19 pandemic and the chronic disease prescription.

2. Methods

Our methodology includes various steps that are consolidated via an open-source[3] implementation in the form of a single Python function. Below is an overview of the workflow followed by the algorithm during its execution

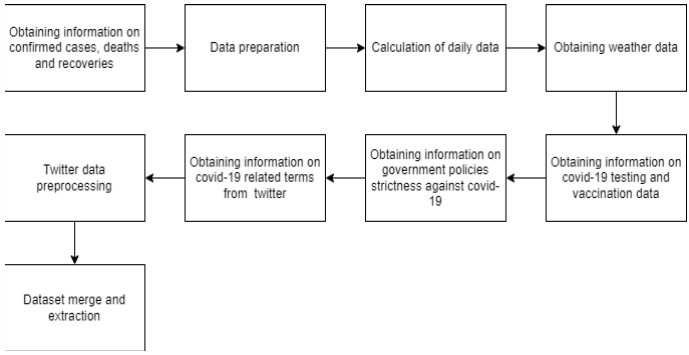


Figure 1. Algorithm execution flow.

When called, the function first retrieves data from a repository [3] about cumulative COVID-19 cases, cumulative recovered and cumulative deaths and transforms the dataset into a time series. Continuing the time series is enriched by calculating the daily changes of the cumulative variables by subtracting the next day from the one before. At this point the nearest weather station to the center of the chosen country is accessed and hourly weather data [4] is retrieved and transformed into daily average values.

The next step in the process is to retrieve information about the selected country's COVID-19 testing, vaccinations and the strictness of the policies adopted [5][6][7]. The stringiness of the policies adopted is rated on a scale from 1 through 5, and includes quantification of policies such as restrictions on gatherings, work from home, facial coverings etc. At this point new variables are calculated combining the ones retrieved as well as preprocessing is used to treat missing values and clean up data. That is achieved either by linear interpolation or corrections of mistakes in the data values that are gathered from the data sources such as duplicates entries for certain dates.

Finally, a plethora of repositories' subdirectories are accessed to retrieve information about daily values of the usage of COVID-19 related terms in twitter [8] and are added to the time series, after which all the collected, created and cleaned up datasets are merged in one which is returned through the function.

[2] The PrescIT project: https://www.prescit.com/
[3] GitHub: https://github.com/gvangelatos/atlantes

The function that implements the aforementioned methodology is entitled "creatCovidDataframe ()". It is written using the Python programming language utilizing several of its available packages. The function when called automatically collects, prepares, creates and consolidates data up to one day before the date it was called.

It is important to note that particular weight has been given to the data processing step, in order to automatically extract data that are immediately available for any further data analysis. During execution the algorithm takes care of any discontinuity in the data, as well as the uniformity of the time intervals between the entries of the data set. Time intervals uniformity is achieved by carefully selecting repositories with daily available data or calculating daily data from repositories that offer a higher sample rate, such as hourly.

3. Results

The function "creatCovidDataframe()" accepts only one argument, which is the country for which data will be exported. At the time of writing the function supports a total of 121 countries which are available on GitHub. During execution the function prints on which part of the algorithm it is, as well as the percentage on which it is currently on as shown in Figure 2.

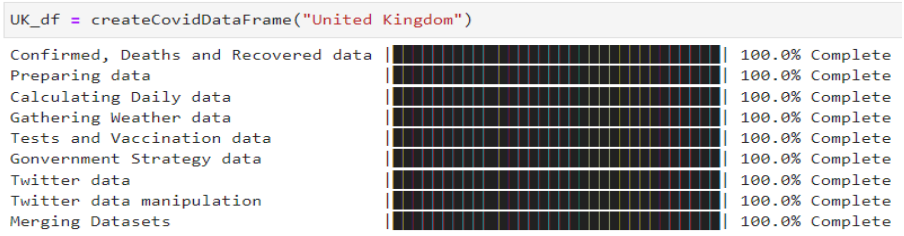


Figure 2. Algorithm runtime output.

The data set exported by the function, consists of daily data, with the first date starting on 01/01/2020 and the last being the day before the one on which it was called. In the last dates in some of the variables that make up the time series, due to a linear interpolation that is applied to the dataset and because their sources do not have new data yet as they are updated every other day, the value of these variables remains constant for 2 and / or 3 days. That is the result of the linear interpolation function applied that carries forwards the data values to fill all the available dates.

In total the function generates a data set which is a time series with sampling time per day and 98 parameters / variables for each day. These 98 variables combine to form the following categories:

- Daily weather data
- Government measures and data on austerity measures
- Global usage terms for COVID-19 on Twitter
- Country statistics
- Daily and cumulative data COVID-19
- COVID-19 Testing and Vaccination data
- Daily admissions to hospitals and intensive care units (some countries)

Comparing our presented methodology with other tools available right now as Open-source solutions, the inclusivity and availability of data are evident. While other solutions

offer only region specific data or provide tracking for only one variable throughout numerous countries, our presented tool is able to access COVID-19 related data for any one of the 121 countries in minutes as shown in Figure 3.

```
--- Maximum observed execution time: 376.6089494228363  seconds ---
--- Minimum observed execution time: 118.51333141326904 seconds ---
--- Average observed execution time: 211.48482256218537 seconds ---
```

Figure 3. Algorithm calculated execution times.

In addition, most solutions for accessing data require extra programming steps in order to access their remote repositories (CSSEGISandData) or they work by manually downloading the datasets (ourworldindata) and integrating them into the code. While some limit the data categories strictly to COVID-19 related some of the datasets are not even in a daily series format. The function presented is designed to fully automate this process accessing, merging and preprocessing data to integrate it into a coherent dataset with predefined units of measurement and time frames.

Finally, in Figure 4 we present two from the many interesting visualizations resulting from variables contained within the exported dataset from "creatCovidDataframe()" showcasing the huge scientific interest that exploration of variables whose impact hasn't yet been accurately measured such as social media and weather data pose.

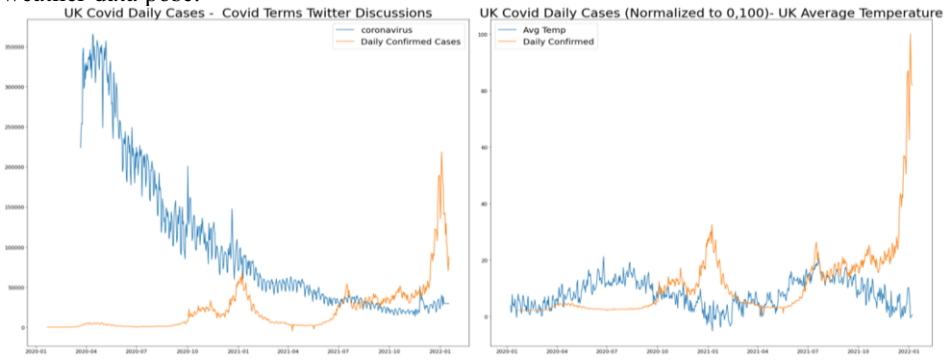


Figure 4. Daily social media and weather data plots.

4. Discussion

As countries experiment with ways to limit the spread of COVID-19, and as analysts continue working on unlocking the secrets of mitigating the spread of the pandemic, they will need more and more data. Thus a way to acquire clean and trustworthy data in a fast and reliable way is of outmost importance in the race against the spread.

Furthermore, understanding the inner workings of the progression of the pandemic proves troublesome while some areas seem to be doing a better job than others adopting the same measures; the reasons behind this still remain undiscovered. As a result a widening of the search for variables explaining that phenomenon is required, something that call for variables like temperatures and twitter metrics being available for the same time frames as the COVID-19 data and being accurately and frequently collected.

5. Conclusion

The data generated by our methodology is largely ready to be used in data analytics and predictive models, helping to mitigate the problems faced by COVID-19 and other data analysts. The implemented function tries to solve the problems of the data being scattered, not designed to be combined, in different time periods and that its volume is constantly increasing by offering a fast, automated and trustworthy solution to acquiring data and removes most of the need for pre-processing.

Finally, within the exported data, are included variables, still unknown whether they do offer any insight to the pandemic, or ones whose impact hasn't yet been accurately measured due to the lack of data. As such, government policies adopted, temperature values, social media metrics and country related statistics are included in the dataset as a possibility to hold the key to further our understanding of the pandemic.

Moving forward, we need to overcome limitations in order to be able to improve Atlantes. Such limitations are finding sources that deliver data daily or switch to sources with an even greater level of credibility. In addition, looking into the future we aspire to incorporate even more variables in the collection of Atlantes such as cases by gender as well as use KB to support is the investigation of relations among the COVID-19 pandemic and the chronic disease prescription by incorporating related data.

Acknowledgement

This research has been co-financed by the European Regional Development Fund of the European Union and Greek national funds through the Operational Program Competitiveness, Entrepreneurship and Innovation, under the call RESEARCH – CREATE – INNOVATE (project code: T2EDK-00640).

References

[1] WHO Director-General's remarks at the media briefing on COVID19- 03/2020 [Google Scholar]
[2] Karanikas H, et al., Development Of Prescription E-Protocols For Medicines And Integration On The Greek National E-Prescription System, Value Heal. 18 (2015) A385. doi:10.1016/j.jval.2015.09.835.
[3] Dong E, Du H, Gardner L. An interactive web-based dashboard to track COVID-19 in real time. Lancet Inf Dis. 20(5):533-534. doi: 10.1016/S1473-3099(20)30120-1
[4] Meteostat: The Weather's Record Keeper (Source: Meteostat)
[5] Mathieu E, Ritchie, H, Ortiz-Ospina E, et al. A global database of COVID-19 vaccinations. Nat Hum Behav (2021). https://doi.org/10.1038/s41562-021-01122-8 .
[6] Hasell J, Mathieu E, Beltekian D, et al. A cross-country database of COVID-19 testing. Sci Data 7, 345 (2020). https://doi.org/10.1038/s41597-020-00688-8
[7] Hale T, Angrist N, Goldszmidt R, et al. A global panel database of pandemic policies (Oxford COVID-19 Government Response Tracker). Nature Human Behaviour. 2021). https://doi.org/10.1038/s41562-021-01079-8
[8] Banda JM, Tekumalla R, Wang G, et al. A large-scale COVID-19 Twitter chatter dataset for open scientific research - an international collaboration [Data set]. (2022). In Epidemiologia (Version 104, Vol. 2, Number 3, pp. 315–324). Zenodo. https://doi.org/10.5281/zenodo.6336166
[9] Painuli D, Mishra D, Bhardwaj S, et al. Forecast and prediction of COVID-19 using machine learning. (2021). Data Science for COVID-19, 381–397. https://doi.org/10.1016/B978-0-12-824536-1.00027-7

Challenges of Trustable AI and Added-Value on Health
B. Séroussi et al. (Eds.)

doi:10.3233/SHTI220552

Discovery of COVID-19 Symptomatic Experience Reported by Twitter Users

Keyuan JIANG[a,1], Minghao ZHU[b] and Gordon R. BERNARD[c]
[a]*Purdue University Northwest, Hammond, Indiana, U.S.A.*
[b]*College of Electronic & Information Engineering, Tongji University, Shanghai, China*
[c]*Vanderbilt University, Nashville, Tennessee, U.S.A.*

Abstract. Since the beginning of the COVID-19 pandemic, patients shared their personal experiences of the viral infection on social media. Gathering their symptomatic experiences reported on Twitter may help better understand the infectious disease and supplement our knowledge of the disease gathered by healthcare workers. In this study, we identified personal experience tweets related to COVID-19 infection using a pre-trained and fine-tuned language model, and annotated the machine-identified tweets in order to extract the information of infection status, symptom concepts, and the days the symptomatic experience occurred. Our result shows that the top 10 most common symptoms mentioned in the collected Twitter data are in line with those published by WHO and CDC. The symptoms along with the day information appear to provide additional insight on how the infection progresses in infected individuals.

Keywords. Novel coronavirus, COVID-19 symptoms, personal health experience, Transformer-based language model, Twitter data

1. Introduction

The outbreak of the novel coronavirus, which occurred in January 2020, has led to the pandemic of COVID-19, declared by the World Health Organization (WHO) on March 11, 2020. Since the outbreak, this highly contagious, life-threatening virus has infected more than 433 million people and caused over 5.9 million deaths around the world, according to WHO [1]. Even with the successful development and wide availability of various COVID-19 vaccines, the world is still struggling to contain the pandemic.

At the onset of the breakout, it became challenging for healthcare workers to accurately diagnose the disease due to the limited understanding and knowledge of the new disease and the similarity of its symptoms to those of flu and common cold. Healthcare givers were learning about the disease while caring and treating the patients, to gain the knowledge about the disease and to discover effective treatments. COVID-19 symptoms were reported by caregivers in the early days of the pandemic [2-4].

Meanwhile, COVID-19 symptoms were shared on social media by those who experienced the viral infection. Twitter posts have received noticeable attention since the beginning of the pandemic and efforts were made to leverage the Twitter data to understand the new outbreak. Chen and colleagues [5] started collecting COVID-19

[1] Corresponding Author, Keyuan Jiang, Department of Computer Information Technology and Graphics, Purdue University Northwest, Hammond, Indiana 46323, U.S.A; E-mail: kjiang@pnw.edu.

related tweets on 28 January 2020. The collection has continued since and is ongoing. Tweet IDs of more than 2 billion tweets collected are publicly shared on GitHub[2]. To make the Twitter data useful for various natural language processing tasks, Müller and colleagues (2020) generated a transformer-based language model, named COVID-Twitter BERT (CT-BERT), based on Google's BERT (Bidirectional Encoder Representations from Transformers) language model [6]. The language model was learned with a corpus of 22.5 million unannotated COVID-19 related tweets. Sarker et al. [7] attempted to investigate self-reported COVID-19 symptoms from Twitter posts, by manually annotating nearly 500,000 tweets. Their study showed that 203 positive-tested Twitter users reported 1,002 symptoms with 668 unique expression.

COVID-19 experiences posted by Twitter users may provide richer information and more details. For example, "*So I tested positive for Covid19 yesterday...[omitted]. Day 1. My Symptoms:. . Throbbing headache with pressure right behind the eyes. Fever ranging from 102-104*" describes (1) whether the infection was confirmed (*tested positive*), (2) the symptoms (*headache, eye pressure,* and *fever*) and (3) when the symptoms were experienced (*Day 1*). Collecting this type of the information can help enhance our knowledge and understanding of the deadly infectious disease. In this study, we attempt to gather such information from the Twitter data, with machine learning methods and manual annotation to discover the mentions of symptoms along with the day information.

2. Method

Our data processing pipeline starts with collecting COVID-19 related Twitter data which are known for noisiness and informal writings. The collected tweets were preprocessed to remove duplicates, retweets (RTs), and non-English tweets. Tweets pertaining to personal experience were identified from the preprocessed tweets using a method based on the fine-tuning of the pre-trained RoBERTa (Robustly Optimized BERT Pretraining Approach developed by FaceBook AI) language model [8]. The personal experience tweets were later processed by MetaMap Lite [9] to identify any potential symptom terms. Afterwards, tweets with any identified symptoms were annotated manually to extract the infection status, day information and symptoms which were missed by MetaMap Lite.

2.1. Identifying Personal Experience Tweets

To predict personal experience tweets (PETs), we utilized our pre-trained and fine-tuned language model based on RoBERTa (Robustly Optimized BERT Pretraining Approach [10]) for identifying PETs related to medication effects [8]. The motivation for this transfer learning is twofold: (1) both problems are somehow similar in the same domain – all about the personal experience in health, and (2) the annotated PETs for COVID-19 symptoms were not available and would take a significant amount of effort to do so.

The RoBERTa-based language model, with a structure of 12 layers, 768 hidden neurons, 12 self-attention heads and 110M parameters, was initially pre-trained with more than 160GB uncompressed texts [10]. The model was further fine-tuned with 12K annotated tweets related to personal experience of medication effects. With the medication effect tweets, the model achieved 0.877 for accuracy, 0.734 for precision,

[2] https://github.com/echen102/COVID-19-TweetIDs

0.775 for recall, and 0.754 for F1 score. The same model was transferred, without relearning/re-training with the COVID-19 data, to identify personal experience tweets related to disease (COIVD-19) symptoms.

2.2. Data

A corpus of 12 million tweets was collected in May 2020 by querying Twitter.com with a home-made crawler which was implemented in compliance with the crawling policy of Twitter.com (as documented in its robots.txt file). The tweets posted between March 11, 2020 and April 23, 2020 were gathered. The keywords used in querying the tweets are: *covid19*, *COVID-19*, *coronavirus*, *Wuhan pneumonia*, and *nCoV*. This corpus of raw tweets was preprocessed, and non-personal experience tweets were filtered out using a pre-trained transformer-based method developed by our team [8]. Afterwards, the personal experience tweets were processed by MetaMap Lite through its RESTFul API. Concepts extracted by MetaMap Lite were considered symptoms if they belong to the semantic type of *sosy* (sign and symptom). Tweets containing no symptom concepts were discarded, and this process yielded about 11K tweets.

Symptomatic experiences occurred on the first 14 days were of interest in this study, as COVID-19 symptoms develop between 1 and 14 days after exposure [13]. Tweets containing day information of experience were extracted using a set of phrases. For instance, for day 1, the following phrases were used: *day 1*, *day1*, *day one*, *1st day*, and *first day*, to cover various ways expressing the first day from the reference point. However, it is noted that this also resulted in tweets containing any days starting with 1, such as day 18. Finally, a corpus of 699 tweets was identified for manual annotation.

2.3. Data Annotation

After gathering the personal experience tweets, we annotated them to extract information of the infection status, symptoms and the days of symptomatic experience. There were two steps of annotation. The first step was to help set the reference points, start points of infection. The labels listed in Table 1 were assigned to tweets during the first step.

Table 1. Labels for the confirmation/status of COVID-19 infection. This information was used to set the reference point.

Label	Description	Tweet Count
Isolation	Isolation, quarantine	93
Confirmed	Tested Positive	15
Infected	Highly certain without testing	372
Suspected	Not sure, without testing	173
Not Personal	Not a personal experience tweet	2
Unrelated	Not related to COVID-19	33
Lockdown	During a lockdown	8
WFH	Work from home	1
Negative	Tested negative	2

The reference point was defined as the day zero (0) of infection. We decided to combine tweets with labels of *isolation*, *confirmed*, *infected* and *suspected* so that their reference points (start points) are the same. The rational was that (1) there was a lack of testing at the time (March and April of 2020), partially due to an increasing demand for testing and lack of available testing kits, (2) those in isolation include individuals who

showed symptoms or had been exposed to the infected individuals, and (3) all of the tweets contain mentions of symptoms.

The second annotation step was to extract symptom expressions, which can be a single word (e.g., *breathlessness*) or made up of multiple words (e.g., *temp at 7 EST is 100.6*). Listed in Table 2 are the top 10 most common symptoms in our data set after the second step of annotation. As shown in the table, there are multiple ways to express each symptom concept. In particular, each of the three symptom concepts listed (Fever, Breath Difficulty, and Lost Smell/Taste) shows more than 60 different expressions.

Table 2. Top 10 most common COVID-19 symptoms. Unique Expression Count: the number of unique expressions for a symptom concept. Count of Mention: the number of mentions of a symptom concept. A single tweet may contain more one mention of the same symptom concept.

Symptom	Unique Expression Count	Count of Mention
Fever	64	314
Cough	23	315
Headache	20	153
Fatigue	10	150
Ache	14	102
Breath Difficulty	64	126
Lost Smell/Taste	67	74
Sore Throat	14	69
Tight Chest	35	56
Chest Pain/Discomfort	25	49

It is also noted that Twitter users use many layman's terms or consumer health vocabulary (CHV) terms to express the symptom concepts. For example, *temp* and *temperature* were commonly used to describe the concept of fever.

3. Result and Discussions

Table 3 below is the result of aligning the mentions of the top 10 most common symptoms with the day information as reported in the Twitter posts. The figure in each cell is the number of mentions of the corresponding symptom.

Table 3. Mentions of top 10 most common COVID-19 symptoms by day. Chest pain symptoms include other chest discomforts other than tight chest.

	Day													
Symptoms	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**
Fever	21	22	30	29	29	13	29	15	17	21	15	11	3	12
Cough	19	26	33	31	36	29	34	17	18	26	15	10	3	16
Headache	10	19	23	13	20	12	21	11	6	16	8	2	6	4
Fatigue	3	10	19	22	19	6	15	16	11	11	6	1	6	1
Ache	11	21	17	10	8	9	6	4	3	7	2	3	0	2
Breath Difficulty	7	14	16	16	16	8	15	6	9	11	11	3	0	3
Lost Smell/Taste	2	2	4	5	15	5	6	3	4	4	1	3	2	2
Sore Throat	15	11	12	5	7	9	5	4	4	0	2	0	1	0
Tight Chest	3	10	6	6	6	1	7	1	0	4	0	0	0	1
Chest Pain	1	3	4	0	6	3	3	4	2	3	0	0	1	0

It is worth noting that the most common symptoms identified from the study Twitter data are well in line with those published by WHO [11] and CDC of the United States [12]. Besides, the symptomatic experiences reported on Twitter provide additional useful information such as the day of a particular symptom and its severity.

The number of mentioning a particular symptom on a particular day may indicate the likelihood that the symptom would be experienced on that day. This information may help us understand how the infection progresses on individuals. It is known that COVID-19 infection starts at the upper respiratory track, and it is interesting to note that the number of mentioning sore throat is the highest on day 1 during the 14 day period, and the number of mentions of other symptoms (fever, cough, headache, fatigue, ache, breath difficulty, and tight chest) increases from day 1. More people reported loss of smell/taste on day 5 than any other days.

4. Conclusion

In this study, we investigated utilizing a pre-trained and fine-turned language model to identify personal experience tweets related to COVID-19 infection. These personal experience tweets were manually annotated for symptoms. Our result shows that the top 10 common symptoms are in line with those reported by both WHO and CDC, and experiences of infected individuals shared online provide additional and more detailed information of the viral infection. This demonstrates the utility of our approach which helps gather additional information and enhances our knowledge of the COVID-19.

References

[1] World Health Organization: Weekly epidemiological update on COVID-19 - 8 March 2022. https://www.who.int/publications/m/item/weekly-epidemiological-update-on-covid-19---8-march-2022.
[2] Chen J, Qi T, Liu L, Ling Y, Qian Z, Li T, Li F, Xu Q, Zhang Y, Xu S, Song Z. Clinical progression of patients with COVID-19 in Shanghai, China. Journal of infection. 2020 May 1;80(5):e1-6.
[3] Guan WJ, Ni ZY, Hu Y, Liang WH, Ou CQ, He JX, Liu L, Shan H, Lei CL, Hui DS, Du B. Clinical characteristics of coronavirus disease 2019 in China. New England journal of medicine. 2020 Apr 30;382(18):1708-20.
[4] Huang C, Wang Y, Li X, Ren L, Zhao J, Hu Y, Zhang L, Fan G, Xu J, Gu X, Cheng Z. Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. The lancet. 2020 Feb 15;395(10223):497-506.
[5] Chen E, Lerman K, Ferrara E. Tracking social media discourse about the covid-19 pandemic: Development of a public coronavirus twitter data set. JMIR public health and surveillance. 2020 May 29;6(2):e19273.
[6] Müller M, Salathé M, Kummervold PE. Covid-twitter-bert: A natural language processing model to analyse covid-19 content on twitter. arXiv preprint arXiv:2005.07503. 2020 May 15.
[7] Sarker A, Lakamana S, Hogg-Bremer W, Xie A, Al-Garadi MA, Yang YC. Self-reported COVID-19 symptoms on Twitter: an analysis and a research resource. Journal of the American Medical Informatics Association. 2020 Aug;27(8):1310-5.
[8] Zhu M, Song Y, Jin G, Jiang K. Identifying personal experience tweets of medication effects using pre-trained Roberta language model and its updating. InProceedings of the 11th international workshop on health text mining and information analysis 2020 Nov (pp. 127-137).
[9] Demner-Fushman D, Rogers WJ, Aronson AR. MetaMap Lite: an evaluation of a new Java implementation of MetaMap. Journal of the American Medical Informatics Association. 2017 Jul 1;24(4):841-4.
[10] Liu Y, Ott M, Goyal N, Du J, Joshi M, Chen D, Levy O, Lewis M, Zettlemoyer L, Stoyanov V. Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692. 2019 Jul 26.
[11] World Health Organization: Common symptoms of COVID-19. https://www.who.int/mongolia/multi-media/item/common-symptoms-of-covid-19
[12] Centers for Disease Control and Prevention. Symptoms of COVID-19. https://www.cdc.gov/coronavirus/2019-ncov/symptoms-testing/symptoms.html
[13] World Health Organization: Diagnostic testing for SARS-CoV-2. https://apps.who.int/iris/bitstream/handle/10665/334254/WHO-2019-nCoV-laboratory-2020.6-eng.pdf.

Challenges of Trustable AI and Added-Value on Health
B. Séroussi et al. (Eds.)

doi:10.3233/SHTI220553

Global Good Open Source Software Development in Response to the COVID-19 Pandemic – Perspectives from SORMAS Implementation in Europe

Candice LOUW[a,1], Rafael PAFFENHOLZ[b], Carolin VERSET[c] and Gérard KRAUSE[a,d,e]
[a] *Helmholtz-Zentrum für Infektionsforschung, HZI, Braunschweig, Germany*
[b] *Gesundheits-, Sozial- und Integrationsdirektion Kantonsärztlicher Dienst, Kanton Bern, Switzerland*
[c] *Agence Régionale de Santé, Bourgogne-Franche-Comté, France*
[d] *Medizinische Hochschule Hannover, MHH, Hannover, Germany*
[e] *Deutches Zentrum für Infektionsforschung, DZIF, Braunschweig, Germany*

Abstract. In recent years, software has evolved from being static, closed source, proprietary products to being dynamic, open source, ecosystems contributing to the global good. To this end, the open source software (OSS) solution and global good, Surveillance Outbreak Response Management and Analysis System (SORMAS), rapidly adjusted to the demands of the Coronavirus disease 2019 (COVID-19) outbreak by introducing a COVID-19 module. This allowed countries that were already making use of the software as part of their public health surveillance infrastructure to make use of the new module in order to respond to the pandemic. New countries in continental Europe, most notably Germany, Switzerland, Liechtenstein and France subsequently chose to adopt the software for public health surveillance purposes for the first time during 2020, requiring additional adaptations to meet local needs. As a result, in this paper, we aim to gain a better understanding of how rapidly SORMAS was adapted to meet global needs by analyzing the SORMAS COVID-19 module introduction timeline, as well as the overall development activity of the software during 2020 and 2021 in response to the pandemic. Favorable initial feature response times in combination with development scale-up possibilities speak to some of the potential advantages of implementing global good OSS tools such as SORMAS for public health surveillance, in response to an emergency. Overall, SORMAS serves as proof of concept for developing a global good OSS solution on an international scale.

Keywords. Open Source Software, Software Ecosystem, Digital Transformation, Health Informatics, COVID-19, Europe

1. Introduction

In recent years, software has evolved from being static, closed source, proprietary products to being dynamic, open source, ecosystems contributing to the global good.

[1] Candice Louw, Helmholtz-Zentrum für Infektionsforschung, Braunschweig, Germany, 38124; E-mail: candice.louw@helmholtz-hzi.de.

The ever changing, independent development of source code and associated components on open source platforms may exhibit a close resemblance to that of a virus – constantly mutating in order to adjust to its environment and host requirements. Recognizing the benefits of implementing open source software (OSS) solutions in the context of public health becomes particularly important in this case, given the flexibility, interoperability and resource savings characteristics that these solutions may offer [1, 2].

To this end, the digital public health OSS solution and global good, Surveillance Outbreak Response Management and Analysis System (SORMAS), rapidly adjusted to the demands of the Coronavirus disease 2019 (COVID-19) outbreak by introducing a COVID-19 disease module [3]. This allowed countries that were already making use of the software as part of their public health surveillance infrastructure (for diseases other than COVID-19) to make use of the new module in order to respond to the pandemic.

Despite SORMAS being primarily programmed in Europe [3], the countries making use of the software at the time of introducing the COVID-19 module were of an international nature (i.e. outside of Europe) [4]. During 2020 however, European countries, including the likes of Germany [3], Switzerland (including Liechtenstein) [5] and France [6], adopted the software for public health purposes for the first time.

Armed with the knowledge that open source global goods can be adapted to meet local needs during a disaster response [7], in this paper, we aim to gain a better understanding of how rapidly SORMAS was adapted to meet not only European, but also international needs. As a result, in this paper, we aim to gain a better understanding of how rapidly SORMAS was adapted to meet global needs by analyzing the SORMAS COVID-19 module introduction timeline, as well as the overall development activity of the software during 2020 and 2021 in response to the pandemic. We delve into the details of the chosen methods next.

2. Methods

From a functional perspective, SORMAS is a public health surveillance tool used to capture data on cases and contacts of over 20 types of infectious diseases. It offers a range of features and functionalities that include (at the time of writing) a dashboard with an overall system statistical overview, task management, person search, case and contact management (including follow-up), event management, sample management and immunization (including vaccination) capturing.

Due to the open source nature of the project, the software is continuously evolving based on user needs, influence and feedback, with a new version of the software being published roughly every 3-6 weeks under the GPL v.3 license. Management and documentation of the SORMAS source code repositories, backlogs and specifications happens on the open source platform GitHub (https://github.com/hzi-braunschweig). In the SORMAS-Docker sub-project (https://github.com/hzi-braunschweig/SORMAS-Docker and https://hub.docker.com/orgs/hzibraunschweig), prefabricated container images are provided that serve as basis for a majority of SORMAS installations.

Our research consists of two main parts including (1) identifying when the COVID-19 module was introduced to the SORMAS source code and (2) the analysis of the source code contributions (commits) to the SORMAS GitHub repository over the project's open source lifetime (till the end of 2021).

For part 1, we identify how rapidly the COVID-19 module was introduced to the software by analyzing the GitHub user story creation date and comparing it to the release

date of the version it was made available in. For part 2, we use git-bars, a utility that employs git log to render simple git commit activity bars on the terminal [8]. Commit activity is group and analyzed by year and month. This gives an overview of the development activities on the repository at two different levels of granularity, allowing identification of average development trends, as well as notable deviations from this. We discuss the obtained results next.

3. Results

3.1. SORMAS COVID-19 Module

The creation of the first COVID-19 module feature request on the SORMAS GitHub repository took place on 2020-01-24 [9]. Implementation of the feature followed and introduction of the new module was on 2020-01-29 as part of the SORMAS version 1.32.0 release [10]. Within a week of specification, the COVID-19 module was thus developed and ready for deployment to SORMAS instances all over the world.

3.2. SORMAS Repository Analysis

We utilized the open source tool "git-bars" in January of 2022 to retrospectively analyze and visualize the SORMAS GitHub repository commit activity. For the timeframe from 2016-05 (date that SORMAS was first published open source on GitHub) to 2021-12, a total of 18534 commits to the project source code were recorded. When expressing this value in terms of years by running the command ">git-bars -p year", the result of 18534 commits, grouped over a 6 year timeframe is delivered, as is visible in Figure 1.

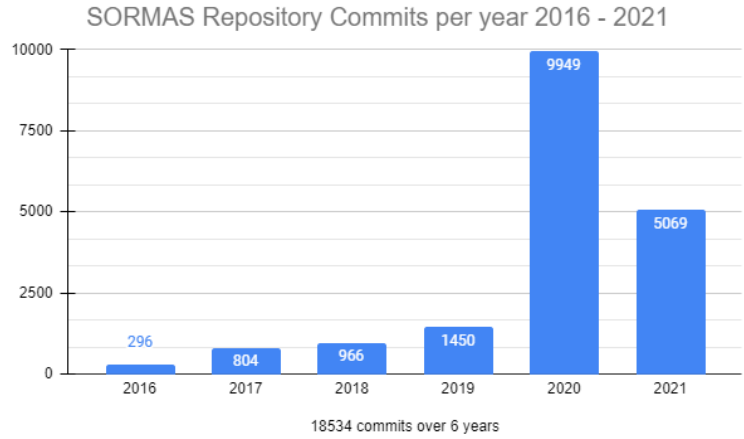


Figure 1. SORMAS GitHub repository commit activity grouped by year (2016 - 2021).

From the results presented in Figure 1, it is possible to see not only the most commits happening in the year 2020, but also the exponential growth during 2020 when compared to previous years. When analyzing the repository on a monthly level by running the command ">git-bars -p month", results indicate 18534 commits over 68 months as summarized in Figure 2.

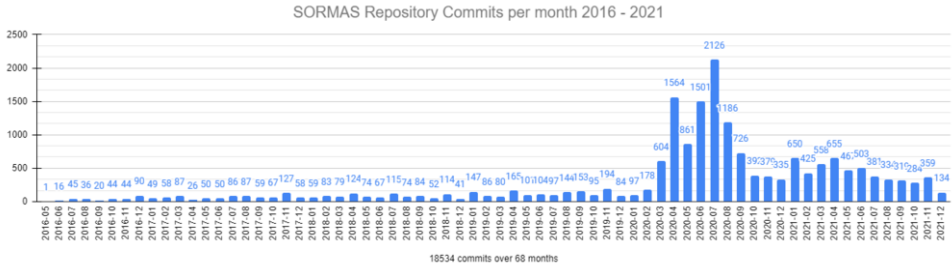


Figure 2. SORMAS GitHub repository commit activity grouped by month (2016 - 2021).

From the results presented in Figure 2, an increase in monthly commit activity is clearly visible in 2020-03, with 604 commits compared to the 178 commits in the preceding month. During the collective preceding timeframe of the software's open source lifetime (2016-05 – 2020-02) commits never exceeded 200 per month. In 2020-03, however, it is possible to observe a substantial increase (threefold) in commits, with commit activity only sinking below the 200 mark again in 2021-12. During the timeframe 2020-03 – 2021-11 (21 months), the overall commit activity of SORMAS was therefore notably higher than previously (2016-05 – 2020-02, 46 months).

4. Discussion

The first cases of COVID-19 were recorded in December 2019 and in March 2020, the World Health Organization (WHO) declared the COVID-19 outbreak a pandemic [11, 12]. SORMAS introduced the COVID-19 disease module by the end of January 2020 making the software ready to respond to the pandemic a month before the WHO officially made its declaration. The favorable new disease module introduction time of under 2 weeks from specification to implementation, thus contributed greatly to establishing the relevance of the software for pandemic response.

The notable increase in SORMAS repository commits, starting in 2020-03, align with the pandemic declaration of the WHO during the same month. It is thus arguable that the increase in commit activity is directly related to the COVID-19 pandemic and subsequent further development of the module in SORMAS. It is, however, important to note that not all commit activity is necessarily only related to the development of the COVID-19 module as SORMAS contains over 40 diseases (aggregate and case-based combined). Despite not being able to say that all commits are for the development of the COVID-19 module exclusively, the introduction of the module most certainly plays a major role in the increased development activity of the software, implying that the COVID-19 pandemic had a direct impact on the increased development activity of the software during 2020 and 2021.

First time adjustment and implementation of the software for use in Germany during Q2 of 2020, as well as Switzerland (including Liechtenstein) and France during Q3 of 2020, certainly also contributed to the above average commit activity during the 2020 timeframe as adjustments were made to the module to accommodate local needs. As a result, by the end of 2021, an estimated 150/375 local health authorities in Germany, 8/18 regions in France, and 13/26 cantons in Switzerland, as well as Liechtenstein, were

either actively using SORMAS, or were in the process of starting to use SORMAS, with at least the COVID-19 module active. Further assessments of SORMAS pertaining to management and in-country use (concerning for example user acceptance, efficiency, data quality, testing etc.), fall outside the scope of this paper and are left to future work.

5. Conclusion

SORMAS serves as proof of concept for developing a global good OSS solution for public health surveillance, and to our knowledge, is one of the first OSS solutions for public health deployed concurrently for usage in multiple European countries (i.e. Germany, Switzerland and Liechtenstein, as well as France during 2020), in addition to other countries around the world. The software demonstrates that open source global goods can rapidly be adapted to meet not only local needs in an emergency [7], but arguably also global needs. Overall, favorable initial feature response time in combination with development scale-up possibilities, speak to the potential advantages of implementing OSS tools such as SORMAS for public health.

References

[1] Neofytou MS, et al. Electronic Health Record Application Support Service Enablers, 2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), 2015, pp. 1401-1404, DOI: https://doi.org/10.1109/EMBC.2015.7318631.
[2] European Commission, 2020, COMMUNICATION TO THE COMMISSION, OPEN SOURCE SOFTWARE STRATEGY 2020 – 2023, Think Open. Available from: https://ec.europa.eu/info/departments/informatics/open-source-software-strategy_en (Accessed 18 December 2021).
[3] SORMAS ÖGD COVID-19. 2021. Kontaktpersonen-management im ÖGD. Available from: https://www.sormas-oegd.de/ (Accessed 19 January 2022).
[4] Helmholtz-Zentrum für Infektionsforschung. 2021. SORMAS®-open, Surveillance, Outbreak Response Management and Analysis System. Available from: https://www.helmholtz-hzi.de/de/forschung/forschungsprojekte/ansicht/projekt/detail/sormasr-open (Accessed 12 January 2022).
[5] Bote.ch. 2020. Contact-Tracing: Bund lässt deutsche Software für 700'000 Franken «helvetisieren». Available from: https://www.bote.ch/nachrichten/schweiz/contact-tracing-bund-laesst-deutsche-software-fuer-700000-franken-helvetisieren;art177490,1257113 (Accessed 25 January 2022).
[6] CGI. 2020. CASE STUDY. Assisting a French regional health agency in operationalizing pandemic measures to save lives. Available from: https://www.cgi.com/en/case-study/reinvent/assisting-a-french-regional-health-agency-in-operationalizing-pandemic-measures-to-save-lives (Accessed 25 January 2022).
[7] Burke W. Mamlin J, Shivers E, et al. OpenMRS as an emergency EMR—How we used a global good to create an emergency EMR in a week, International Journal of Medical Informatics, Volume 149, 2021, 104433, ISSN 1386-5056, DOI: https://doi.org/10.1016/j.ijmedinf.2021.104433.
[8] Git-bars. 2020. Available from: https://github.com/knadh/git-bars (Accessed 28 January 2022).
[9] GitHub issue. 2020. Coronavirus [2] #1476. Available from: https://github.com/hzi-braunschweig/SORMAS-Project/issues/1476 (Accessed 25 January 2022).
[10] GitHub release. 2020. SORMAS 1.32.0. Available from: https://github.com/hzi-braunschweig/SORMAS-Project/releases/tag/v1.32.0 (Accessed 25 January 2022).
[11] World Health Organization Regional Office for Europe. 2020. WHO announces COVID-19 outbreak a pandemic. Available from: https://www.euro.who.int/en/health-topics/health-emergencies/coronavirus-covid-19/news/news/2020/3/who-announces-covid-19-outbreak-a-pandemic (Accessed 25 January 2022).
[12] World Health Organization. 2020. Archived: WHO Timeline - COVID-19. Available from: https://www.who.int/news/item/27-04-2020-who-timeline---covid-19 (Accessed 25 January 2022).

Challenges of Trustable AI and Added-Value on Health
B. Séroussi et al. (Eds.)

doi:10.3233/SHTI220554

The COVID-19 Data Exchange Platform of the German University Medicine

Hans-Ulrich PROKOSCH[a,1], Thomas BAHLS[b], Martin BIALKE[b], Jürgen EILS[c], Christian FEGELER[d], Julian GRUENDNER[a], Birger HAARBRANDT[e], Christopher Hampf[b], Wolfgang HOFFMANN[b], Hauke HUND[d], Marvin KAMPF[f], Lorenz A KAPSNER[f], Piotr KASPRZAK[n], Oliver KOHLBACHER[g,l,m], Dagmar KREFTING[h], Jonathan M MANG[f], Michael MARSCHOLLEK[e], Sebastian MATE[f], Armin MÜLLER[c], Fabian PRASSER[c], Julian SASS[c], Sebastian SEMLER[i], Holger STENZHORN[g,j], Sylvia THUN[c], Sven ZENKER[k] and Roland EILS[c]

[a]*Chair of Medical Informatics, Friedrich-Alexander-Universität Erlangen-Nürnberg, Erlangen, Germany*
[b]*Institute for Community Medicine, University Medicine Greifswald, Greifswald, Germany*
[c]*Berlin Institute of Health at Charité – Universitätsmedizin Berlin, Berlin, Germany*
dGECKO Institute, Heilbronn University of Applied Sciences, Heilbronn, Germany
[e]*Peter L. Reichertz Institute for Medical Informatics of TU Braunschweig and Hannover Medical School, Hannover, Germany*
[f]*Medical Center for Information and Communication Technology, Universitätsklinikum Erlangen, Erlangen, Germany*
[g]*Institute for Translational Bioinformatics, University Medical Center, Tübingen, Germany*
[h]*Department of Medical Informatics, University Medical Center Göttingen, Göttingen, Germany*
[i]*TMF - Technology, Methods, and Infrastructure for Networked Medical Research, Berlin, Germany*
[j]*Institute for Medical Biometry, Epidemiology und Medical Informatics, Saarland University Medical Center, Homburg, Germany*
[k]*Staff Unit for Scientific & Medical Technology Development & Coordination (MWTek), Commercial Directorate; Institute for Medical Biometry, Informatics & Epidemiology; Department of Anesthesiology and Intensive Care Medicine, University Hospital Bonn, Bonn, Germany*
[l]*Institute for Bioinformatics and Medical Informatics, University of Tübingen, Tübingen, Germany*
[m]*Department of Computer Science, University of Tübingen, Tübingen, Germany*
[n]*Gesellschaft für wissenschaftliche Datenverarbeitung mbH, Göttingen, Germany*

Abstract. COVID-19 has challenged the healthcare systems worldwide. To quickly identify successful diagnostic and therapeutic approaches large data sharing approaches are inevitable. Though organizational clinical data are abundant, many of

[1] Corresponding Author, Hans-Ulrich Prokosch, Chair of Medical Informatics, Friedrich-Alexander-Universität Erlangen-Nürnberg, Wetterkreuz 15, 91058 Erlangen, Germany; E-mail: Hans-Ulrich.Prokosch@fau.de.

them are available only in isolated silos and largely inaccessible to external researchers. To overcome and tackle this challenge the university medicine network (comprising all 36 German university hospitals) has been founded in April 2020 to coordinate COVID-19 action plans, diagnostic and therapeutic strategies and collaborative research activities. 13 projects were initiated from which the CODEX project, aiming at the development of a Germany-wide Covid-19 Data Exchange Platform, is presented in this publication. We illustrate the conceptual design, the stepwise development and deployment, first results and the current status.

Keywords. Real world data, data sharing network, Covid-19, pandemic preparedness

1. Introduction

Real world data analysis in medicine today relies on the availability of clinical data collected during clinical care processes [1]. The harnessing and cross-consortial use of such data is one of the major goals of the Federal Ministry of Education and Research (BMBF)-funded Medical Informatics Initiative (MII) and its four consortia [2]. The collection, exchange and joint analysis of diverse data referring to COVID-19 patients has been an important method of the network university medicine (NUM) that has been founded in April 2020 to understand and manage the COVID-19 pandemic in Germany. As one NUM project, the COVID-19 Data Exchange Platform CODEX has been designed and implemented by the MII partners to enable all university hospitals in Germany to harmonize, share and analyze COVID-19 real world data.

The objective of this publication is to describe the project´s conceptual approach, the stepwise development and deployment, first results and the current CODEX status.

2. Methods

To quickly initiate the CODEX project [3] most of the German university hospitals built on the data integration centers, governance, and methods established within the MII. CODEX brings together the competencies and previous work of more than 15 university hospitals and industrial partners making the results of their work available to the scientific community as open-source software. Further design principles were vendor neutrality and open application programming interfaces.

A first coordinated step within NUM was the interdisciplinary definition and specification of the FHIR-based GECCO (German Corona Consensus Dataset) dataset [4], with data elements and response options being semantically mapped to e.g., SNOMED CT and LOINC for international comparability. To quickly provide a harmonized deployable data repository environment we started with distributed i2b2 repositories, which had already proven their capabilities within MIRACUM [5] and a cross-consortial MII demonstrator study [6]. It was also decided to further build on the already established MII concepts, with federated data collection within routine clinical care environments, thus providing federated analysis options, but extend this with a central platform (similar to the N3C collaborative [8]; based on results from the MII HiGHmed consortium [7]). This enabled linkage with for example COVID-19 data from citizen apps. The implementation of a federated Trusted Third Party (fTTP) [9] enables cross-institutional as well as privacy-preserving record linkage (PPRL) and assures

compliance with the requirements of the European General Data Protection Regulation (GDPR). Ethics approval for data transfer to the central CODEX platform was given amongst others by the Erlangen University Ethics Committee (No. 500_20 B; R. Maas).

3. Results

3.1 The decentral CODEX nodes

Local data integration centers (DICs) at each of the university hospitals provide core data infrastructure for routine clinical care data extraction, harmonization, and integration. Based on these, CODEX implemented 'CODEX nodes' – harmonized local software infrastructures for providing the GECCO data. The nodes rely on a set of centrally provided software components, which had to be combined and integrated into the local IT infrastructures. The core components comprise:

- an **i2b2 repository** with patient demographics, diagnosis, and procedures as well as ETL processes to fill it from standardized German billing data records
- a **FHIR server** to provide the GECCO dataset
- the **Trusted Third Party components** to support local pseudonymization (gPAS®), record linkage (E-PIX®), consent management (gICS®), local provisioning of standardized FHIR consent resources and support for a FHIR-based linkage to the federated Trusted Third Party (fTTP) [9]
- the **'feasibility triangle'**, named after its three components, providing (1) a FHIR Server, (2) FHIR Search, respectively CQL, execution environments and (3) a middleware for secure linkage with the central feasibility user interface
- the **FHIR and business process engine** (BPE) servers of the HiGHmed Data Sharing Framework (DSF) [10] enable secure transport of GECCO datasets between local nodes and the central platform.
- an **EDC system** (e.g., REDCap [11]) to capture GECCO data which were not yet part of the electronic documentation at the German University Hospitals
- the **ODM2FHIR converter** to transform standardized EDC system exports in ODM format to the GECCO FHIR profile.

These components were developed, adapted, and pre-configured by consortium members and provided to all partners within the network together with templates for the required documentation (e.g., data protection concepts, IT security documentation) to accelerate its use in the highly regulated clinical environment. The development and deployment of these components were tightly coordinated through a series of weekly video conferences.

3.2 The central CODEX platform components

The central CODEX platform extends the federated analysis capabilities of the CODEX nodes, with the possibility of linking patient data across institutions and providing data usage for specific research projects. This is implemented by the following components:

- **the** central **research data platform** is based on EHRbase, an open-source software backend for electronic health records based on OpenEHR [12]. Access to

the platform is provided through a central portal enabling cohort selection and data export requests to answer specific research questions.

- **the DSF-**based **GECCO Transfer Hub** acts as a gatekeeper and orchestrates the transfer of data from the CODEX nodes to the central research data platform including a pseudonym exchange using the fTTP.
- **a** federated **Trusted Third Party** (fTTP) performs privacy-preserving record linkage and pseudonymization for the central platform during the data transfer process. The functionality is implemented using the open-source tools of the Trusted Third Party Greifswald [9].
- **a CODEX Dashboard** continuously provides anonymized information on the total number of patients with COVID-19 infection, level of care, length of stay in intensive care as well as age and gender [13]. Most recently, a MII core dataset FHIR specification compliant open-source processing engine to feed the dashboard from local FHIR stores has been provided.

A central **Data Use & Access Committee** reviews requests for data usage.

3.3 First CODEX-based research results

Based on the i2b2 deployments provided in early summer 2020 first research questions could be answered. These were for example:

- How did university hospital patient numbers change during the first German COVID-19 lockdown (data from 14 German university hospitals)? [14]
- What was the lethality of COVID-19 patients in the first half year of the pandemic in association with clinical risk factors (data from 18 German university hospitals)? [15]
- Was there an observable shift of radiotherapy use during the first wave of the COVID-19 pandemic (data from 14 German university hospitals) [16]?

Some of the university hospitals leveraged their i2b2 implementations and participated in the international consortium (4CE) of 96 hospitals across five countries (www.covidclinical.net) which aimed to use the largely untapped resource of EHR data to address critical clinical questions about COVID-19 [17].

4. Discussion and outlook

During a short project period a very comprehensive data sharing platform was established across all German university hospitals. The technical implementation, despite its complexity and the widely distributed developer team, could be smoothly established. Towards this end, the project nicely illustrates how integrating the different competencies from the whole university hospital landscape of Germany leads to synergies for an efficient joint data sharing platform development. However, it also illustrated the socio-technical hurdles within bureaucratic environments. The regulatory aspects (e.g., data provisioning contracts between the central platform and all university hospitals, obtaining positive ethics votes for the MII broad consent text extended with the data transfer aspects to the CODEX central platform), proved to be much more challenging. Thus, we could, by the end of 2021, provide more than 17,000 GECCO records within the distributed DIC repositories. To enable their research use local data

use and access committees had authorized provision of their data for the above mentioned federated data analysis projects. Positive ethics votes for transferring data to the central platform, in many hospitals however, could only be obtained very late in the project. Thus, the consent for central data transfer currently only exists for about 350 patients.

Based on the collaboration culture and joint infrastructure established in CODEX in 2021, two follow up projects have been initiated in 2022. The CODEX Routine Data Platform will ascertain sustainability of the established infrastructure by addressing both governance and maintenance tasks and providing deeper integration with MII structures. Starting in 2023 it will be extended to cover the complete MII core dataset. The CODEX+ project extends the achieved results in various directions, including interfacing more deeply with infrastructures previously developed in MII use cases, broadening the data to be included in the platform and linking with international projects.

References

[1] Maissenhaelter BE, Woolmore AL, Schlag PM. Real-world evidence research based on big data: Motivation-challenges-success factors. Onkologe (Berl). 2018;24 (Suppl 2): 91-98.
[2] Semler SC, Wissing F, Heyder R. German Medical Informatics Initiative. Methods Inf Med. 2018;57(S 01):e50-e56.
[3] CODEX | COVID-19 Data Exchange Platform, Netzwerk Universitätsmedizin. https://www.netzwerk-universitaetsmedizin.de/projekte/codex [last visited on January 3rd, 2022]
[4] Sass J, Bartschke A, Lehne M, et al. The German Corona Consensus Dataset (GECCO): a standardized dataset for COVID-19 research in university medicine and beyond. BMC Med Inform Decis Mak. 2020 Dec 21;20(1):341.
[5] Prokosch HU, Acker T, Bernarding J, et al. MIRACUM: Medical Informatics in Research and Care in University Medicine. Methods Inf Med. 2018;57(S 01):e82-e91.
[6] Ganslandt C, Schaaf J, Schepers J, et al. Experiences from the National Demonstrator Study within the German Medical Informatics Initiative. AMIA Conference, Washington, 2019.
[7] Haarbrandt B, Schreiweis B, Rey S, et al. HiGHmed - An Open Platform Approach to Enhance Care and Research across Institutional Boundaries. Methods Inf Med. 2018;57(S 01):e66-e81.
[8] Haendel MA, Chute CG, Bennett TD, et al. The National COVID Cohort Collaborative (N3C): Rationale, design, infrastructure, and deployment. J Am Med Inform Assoc. 2021;28(3):427-443.
[9] Bialke M, Penndorf P, Wegner T, et al. A workflow-driven approach to integrate generic software modules in a Trusted Third Party. J Transl Med. 2015;13:176.
[10] Hund H, Wettstein R, Heidt CM, Fegeler C. Executing Distributed Healthcare and Research Processes - The HiGHmed Data Sharing Framework. Stud Health Technol Inform. 2021 May 24;278:126-133.
[11] https://projectredcap.org/software/ [last visited on January 19th, 2022].
[12] https://openehr.org/ [last visited on January 19th, 2022].
[13] https://coronadashboard.ukbonn.de/ [last visited on January 19th, 2022].
[14] Kapsner LA, Kampf MO, Seuchter SA, et al. Reduced Rate of Inpatient Hospital Admissions in 18 German University Hospitals During the COVID-19 Lockdown.Front. Public Health. 2020;8:59411.
[15] Schüttler J, Mang JM, Kapsner LA, et al. Letalität von Patienten mit COVID-19: Untersuchungen zu Ursachen und Dynamik an deutschen Universitätsklinika. Anästh. Intensivmed 2021; 62:244–257.
[16] Medenwald D, Brunner T, Christiansen H, et al. Shift of radiotherapy use during the first wave of the COVID-19 pandemic? An analysis of German inpatient data. Strahlenther Onkol (accepted).
[17] Brat GA, Weber GM, Gehlenborg N et al. International electronic health record-derived COVID-19 clinical course profiles: the 4CE consortium. npj Digi. Med 3, 109, 2020.

Challenges of Trustable AI and Added-Value on Health
B. Séroussi et al. (Eds.)

doi:10.3233/SHTI220555

Preliminary Validation of a Rule-Based System for Mortality Coding Using ICD-11

Mihai Horia POPESCU[a], Can CELIK[b], Vincenzo DELLA MEA[a] and Robert JAKOB[b]
[a] *University of Udine, Udine, Italy*
[b] *World Health Organization, Geneva, Switzerland*

Abstract. A crucial process for world-level mortality statistics is the capability to identify the underlying cause of death from death certificates. Currently such certificates are coded using ICD-10. The selection of the underlying cause is done by means of semi-automated rule-based systems. However, starting from 2022, countries should begin to adopt ICD-11, for which no system is already available. The present paper describes the architecture of a novel system for automated UC selection, with classification-independent rules, and its preliminary validation on two sets of death certificates coded with ICD-10 and ICD-11.

Keywords. ICD-11, Cause of Death, medical classifications

1. Introduction

Comparison of mortality statistics is generally done by age and sex, but the so-called underlying cause of death (UC) is the most important information used for such comparison. This is defined by the WHO (2010) as *"I (a) the disease or injury which initiated the train of morbid events leading directly to death; or (b) the circumstances of the accident or violence which produced the fatal injury."* [1]. For each death, the UC is selected from the causal chain of events reported by a physician on the death certificate through the application of specific rules. The death certificates are collected and coded according to a standard methodology defined by the World Health Organisation (WHO) in line with the International Statistical Classification of Diseases and Related Health Problems (ICD) [1,2].

In many nations throughout the world, the coding of the death certificate conditions and the selection of the UC is still done manually. Automated coding systems, on the other hand, have been available since the 1970s, and an increasing number of nations are willing to transition from manual to automated coding.

The available automated systems that are supporting the UC selection are mainly Iris [3,4] and ACME (Automatic Classification of Medical Entry) [5,6] which support the ICD-10 classification, while for ICD-11 there are still no available systems.

ICD has been for more than a century the main basis for comparable statistics on causes of death and non-fatal disease. The 10th revision (ICD-10) was released nearly 30 years ago, which served a variety of functions in at least 120 countries and it has been translated into 43 languages. The 11th revision was adopted by the 72nd World Health Assembly in May 2019 [2]. ICD-11 is not just an extension of categories compared to ICD-10: is a different and more powerful health information system, implemented in modern information technology infrastructures, and flexible enough for future

modification and use with other classifications and terminologies [7]. Instead of the books that represented the official release until ICD-10, ICD-11 is released in form of technological tools like ICD-11 API (Application Program Interface) [8], the Coding Tool [9], and the ICD Field Implementation Tool [10].

In the crucial transition from ICD-10 to ICD-11, any UC selection system will initially suffer of a lack of available datasets already coded with ICD-11. On the other side, the abstract selection algorithm is almost the same for both classifications.

The present paper describes the architecture of a novel system for automated UC selection, with classification-independent rules, and its preliminary validation on two sets of death certificates coded with ICD-10 and ICD-11.

2. Methods

Since the abstract selection algorithm is almost the same for ICD-10 and ICD-11, with differences related to the concrete codes involved, the main requirement for the proposed system is to be classification independent, which means, have a way to separate the selection rules from the actual codes involved. A secondary yet important requirement is the possibility to integrate with the current ICD-11 platform and tools, which in turn are designed for easy integration with third party software. Finally, rules should be easily editable by domain experts.

For the rule-based system we identified two separate modules, one for the implementation of the rule engine, and one for the implementation of the code sets, which in turn could be based on ICD-10, ICD-11 or even an ICD-10 subset called the Start-Up Mortality List (ICD-10-SMoL). The rule engine is implementing the algorithm described in the reference guide [2] described in the sections 2.19-2.20. Selecting the underlying cause of death involves two separate steps. First is it needed to identify the starting point of the sequence of conditions, then to modify the starting point, if any of the modification instructions apply. An example of rule is:

Do not accept Angina pectoris (BA40) and Chronic ischaemic heart disease (BA50-BA5Z) as due to a neoplasm.

The rules format is quite complex, but a simplified version can be viewed as:

```
NAME: Rejected Sequences - Certain ischaemic heart disease due to other
   condition
CONDITION 1: "Certain ischaemic heart disease"
BINARY OPERATION MATCH: "DUE TO"
CONDITION 2: "Neoplasm"
SELECT: "CONDITION 1"
```

This rule is used to select the new tentative UC when Angina pectoris or a Chronic ischaemic heart disease condition it is found to be due to a Neoplasm, and the selected UC is from the condition 1, which is the Certain ischaemic heart disease.

Differently from the other mortality coding systems, that treat codes as “ranges”, described only in terms of leaves of the hierarchical tree, we want to exploit the hierarchy to express the code sets at the highest abstraction level possible.

An example of code set is:

Chronic ischaemic heart disease: which has the range of *BA50-BA5Z* but can be specified in the code set as id *“http://id.who.int/icd/entity/1221742343”* that include all the Chronic ischaemic heart disease conditions.

These are the definitions (not in their JSON syntax for the sake of brevity), of the code sets for ICD-10 and ICD-11 used in the rule above, where they are mentioned by name and not by code:

```
ICD-10: "Angina pectoris" Include "I20", "Chronic ischaemic heart
   disease" Include "I25", "Neoplasm" Include "II", "Certain ischaemic
   heart disease" Include Categories ("Angina pectoris", "Chronic
   ischaemic heart disease").
ICD-11: "Angina pectoris" Include
   "http://id.who.int/icd/entity/718946808", "Chronic ischaemic heart
   disease" Include "http://id.who.int/icd/entity/1221742343", "Neoplasm"
   Include "http://id.who.int/icd/entity/1630407678", "Certain ischaemic
   heart disease" Include Categories ("Angina pectoris", "Chronic
   ischaemic heart disease").
```

2.1. Validation

As already mentioned, validation is an issue because of the lack of certificates dataset coded in ICD-11. However, our classification independent approach allows at least to validate the abstract rules on ICD-10 coded certificates. Thus, we based our validation on a dataset of death certificates coded in ICD-10 from the Center for Disease Control and Prevention (CDC) for the year 2018, plus a small dataset of certificates manually coded in ICD-11, developed ad-hoc for this work. In both datasets, the UC predicted by the system has been compared with the ground truth.

Accuracy has been then calculated, which can be computed as:

$$Accuracy = \frac{Number\ of\ correct\ predictions}{Total\ number\ of\ predictions} \quad (1)$$

3. Results

3.1. The system

A library has been developed to implement the rule-based system, where the related code set has been partially implemented for two ICD classifications (ICD-11 and ICD-10) and ICD-10-SMoL. The library was developed in *.dotnet* framework, which can be deployed in most of the environments. To make use of the library, two applications have been developed: a console-based and a web-based application. The web application back-end was implemented with support of extended technologies like Application Program Interface (API) which gives the opportunity to support implementation of different devices applications.

The rule-based system is implemented in two separate modules, one which specifies the rules in a JSON format, and an algorithm implemented in *C#* that interprets the rules and execute them. Since the rules are defined separately, give the freedom to modify the algorithm logic without changing the associated programming code. This is also the reason why the actual rules do not have any knowledge of the classification behind, but just implement the logic of the algorithm. In order to extend the system with a new classification, a few classes would be need to be extended in the library to support some classification functionalities and implement the related code sets.

The code sets are implemented in multiple JSON files, which the rules refer to for its evaluation, reason why the same code set should be implemented for each classification.

Currently the system fully implements the algorithm, while the code set is partially implemented with similar percentage of completeness for ICD-10 and ICD-11. 18 out of 38 selection rules are fully implemented, where the others need further domain expert support. Near to 95% of the modification rules are implemented, where the remaining need again expert intervention.

Table 1. Effectiveness scores for the analysis of the proposed system, for the analysis made in the Netherlands and for ML approach.

No. certificates	Certificates dataset	System used to select UC	Rejected certificates	Accuracy
2.846.305	ICD-10 dataset	Proposed solution	8.2%	78%
1.248	ICD-11 dataset	Proposed solution	11.6%	62.8%
134.262	ICD-11 Netherlands	Iris [11]	31.5%	78%
400.000	ICD-10 CDC	ML [12]	0%	98.75%

In Table 1 we can observe the effectiveness scores that the proposed system obtained for the ICD-10 dataset and for the ICD-11 dataset. Certificates may be rejected for wrong codes, classification version mismatch, ad in the case of ICD-11, wrong postcoordination. Removing the rejected certificates from the analysis, the system was able to correctly select the UC with an accuracy of 78%. On the ICD-11 dataset, accuracy is 62.8%. In the latter case, the rejected certificates mostly depend on a version mismatch, due to the rapid evolution of ICD-11.

4. Discussion

The preliminary validation of the proposed mortality coding system has been carried out on a large dataset of ICD-10 coded death certificates and on a very small dataset of ICD-11 coded certificates. Although not directly comparable due to the different dataset, for ICD-10 coded certificates preliminary results show an accuracy very close to the one found in a similar study on IRIS [11], while for the ICD-11 data set accuracy is still lower. However, the ICD-11 dataset, which only contains 1248 certificates, might not fully represent the real-world distribution of causes of death, over-representing less frequent cases. On the other side, such accuracy has been obtained with an incomplete set of rules.

A consideration should be done also for systems based on Machine Learning techniques, which have been show to outperform rule-based systems on ICD-10 coded certificates reaching an accuracy of 98.75% [12,13]. Although the results reached, those systems cannot be implemented yet for the new 11th revision. ML has great capabilities and can give great support for this purpose, but to perform they need great quantity of data for the training and the dataset need to be of highest quality, which is not always possible to ensure in an early implementation.

5. Conclusions

In this paper we have presented the first rule-based system with the capability to select the underlying cause of death for the ICD 11th revision, while still being able to work

with the previous revision of the classification. The results seem very promising, and this gives the possibility to mortality coding in the early stage of ICD-11 adoption for statistics, while looking forward for the full implementation of the rules and code sets, and enable a complete validation. A crucial future work, needed for validating this system as well as other future systems, is the development of a data set of ICD-11 coded death certificates, possibly developed by coders in different countries, to cover the inter-country variability in cause of death distribution and coding style.

6. Acknowledgements

We thank the Mexican WHO-FIC Collaborating Center for having developed the ICD-11 coded data set. We thank Robert N. Anderson at the Mortality Statistics Branch, U.S. National Center for Health Statistics, for having provided the ICD-10 coded data set.

References

[1] World Health Organization - WHO. International Statistical Classification of Diseases and Related Health Problems. 10th Revision, vol 2, Fifth edition 2016. Geneva, Switzerland: WHO. Retrieved from https://icd.who.int/browse10/Content/statichtml/ICD10Volume2_en_2016.pdf
[2] World Health Organization. ICD-11 revision. https://icd.who.int/en. Accessed 13 Gen 2022
[3] Iris Institute. Official website. www.iris-institute.org
[4] Pavillon G, Johansson LA, Glenn D, Weber S, Witting B, and Notzon S. 2007. Iris: A Language Independent Coding System For Mortality Data. In WHO-FIC. Annual Meeting. Trieste, Italy, 2007.
[5] Centre for Disease Control and Prevention - CDC, National Centre for Health Statistics – NCHS. National Vital Statistics System. 2021. Instruction manuals. ICD-10 ACME Decision Tables for Classifying Underlying Causes of Death.
[6] Lu TH. Using ACME (Automatic Classification of Medical Entry) software to monitor and improve the quality of cause of death statistics. J Epidemiol Community Health. 2003 Jun; 57(6): 470–471.
[7] Harrison JE, Weber S, Jakob R, Chute CG. ICD-11: an international classification of diseases for the twenty-first century. BMC Med Inform Decis Mak 21, 206 (2021).
[8] Mugisha M, Byiringiro JB, Uwase M, Abizeyimana T, Ndikubwimana B, Karema N, Kayiganwa M, Tran Ngoc C, Kostanjsek N, Celik C, Marc D, Twizere C, Tumusiime D. Integration of International Classification of Diseases Version 11 Application Program Interface (API) in the Rwandan Electronic Medical Records (openMRS): Findings from Two District Hospitals in Rwanda. Stud Health Technol Inform. 2020 Jun 26;272:280-283.
[9] ICD-11 Coding Tool. Mortality and Morbidity Statistics (MMS). https://icd.who.int/ct11/icd11_mms/en/release Accessed 13 Gen 2022
[10] Donada M, Kostanjsek N, Della Mea V, Celik C, Jakob R. Piloting a Collaborative Web-Based System for Testing ICD-11. Stud Health Technol Inform. 2017;235:466-470.
[11] Harteloh P. The implementation of an automated coding system for cause-of-death statistics. Informatics for Health and Social Care. vol 45, no 1, pp 1-14. 2020.
[12] Della Mea V, Popescu MH, Roitero K. Underlying Cause of Death Identification from Death Certificates using Reverse Coding to Text and a NLP Based Deep Learning Approach. Informatics in Medicine Unlocked, 21, 100456 (2020)
[13] Falissard L, Morgand C, Roussel S, Imbaud C, Ghosn W, Bounebache K, Rey G. A Deep Artificial Neural Network-Based Model for Prediction of Underlying Cause of Death From Death Certificates: Algorithm Development and Validation. JMIR Med Inform 2020;8(4):e17125

Challenges of Trustable AI and Added-Value on Health
B. Séroussi et al. (Eds.)

doi:10.3233/SHTI220556

The Impact of Hospital Accessibility on Interregional Patient Mobility in Italy

Fabrizio PECORARO[a,1], Daniela LUZI[a] and Fabrizio CLEMENTE[b]
[a]*Institute for Research on Population and Social Policies, National Research Council. Rome, Italy*
[b]*Institute of Crystallography, National Research Council. Monterotondo (Rome), Italy*

Abstract. Patient mobility represents a proxy measure to assess the quality and availability of hospital services, especially in decentralized health systems. Different studies have been focused on the interregional mobility in Italy to capture factors influencing this phenomenon. Among them, hospital capacity is generally captured through the number of beds per population. However, this indicator does not consider the distance to hospitals and the accessibility of extra-regional beds, in particular for patients living at the regional borders. The aim of this paper is to analyse the effect of extra-regional spatial accessibility component on patient mobility among the Italian regions. This can help to capture the level of equity in the provision of services across the country providing a snapshot of the distribution of beds over the territory. Moreover, this study contributes to gain a deeper understanding of the allocation of health resources providing input for policy makers on the basis of the principles of service accessibility.

Keywords. Spatial accessibility, rehabilitative services, Italy, hospital mobility

1. Introduction

Patient mobility is a complex phenomenon considered as a proxy for the quality and availability of hospital services [1]. This is particularly evident in Italy, a decentralized tax-funded health system affected by significant socio-economic disparities at regional level [2,3]. Moreover, compared with other European countries, in Italy, patients tend more frequently to travel long distances to access to care [4] especially for elective treatments [5]. Patient mobility across Italian regions has been widely studied to capture factors that may influence the patients' choice [1], including social, demographic and economic status [6], quality and complexity of regional services [7] as well as structural components related to personnel, technologies and equipment available [8]. Usually, hospital capacity is assessed analysing the number of beds per population, computed at regional level. However, this indicator considers only the availability of regional resources, neglecting two fundamental aspects of universal care: the accessibility in terms of travel distance and the availability of extra-regional facilities in particular for patients living at the regional borders [9]. Within this context, the aim of this study is twofold. Firstly, it provides an analysis of the extra-regional component of the hospital spatial accessibility outlining a snapshot of the distribution of beds to capture the level of equity across the country [10]. Secondly, it explores the impact of accessing to extra-

[1] Fabrizio Pecoraro, IRPPS-CNR, Via Palestro, 32, 00185 Rome, Italy (f.pecoraro@irpps.cnr.it)

regional resources on patient's mobility. This analysis may provide an input for policy makers to capture to what extent the capacity and distribution of hospitals and beds may affect patient's flows at regional level.

2. Materials and methods

Data on mobility is gathered from the report on hospital admissions [11] published by the Ministry of Health which provides information on activities of public and private hospitals. In this paper we focus the attention on hospitalizations within the rehabilitative wards that provide inpatient services to mitigate or remove orthopaedic and/or neurological issues. These activities are mainly elective treatments where patients are generally prepared to travel beyond their nearest provider in particular in countries that allow patients to freely choose their place of care [12]. This is also confirmed by the high percentage (15.9%) of passive mobility across Italian regions [11]. For each region *i*, a passive mobility index (PM_i) was computed as the rate of hospital discharges of residents occurred in other regions. PM_i can be decomposed as follows to capture the mobility towards each specific region *k* (PM_i^k):

$$PM_i = \sum_{k\in\{k\neq i\}} PM_i^k = \frac{\sum_{k\in\{k\neq i\}} d_i^k}{d_i} = \frac{\sum_{k\in\{k\neq i\}} d_i^k}{\sum_k d_i^k}$$

where d_i^k represents the number of patients residing in region *i* and discharged in region *k* and d_i is the total number of discharges of patients residing in region *i*.

The accessibility index (*AI*) was computed adopting the enhanced two step floating catchment area methodology (E2SFCA) [11]. This method is based on a gravity model which relates the increasing probability to access to a hospital with the number of beds and patient-to-hospital distance. It is calculated for each municipality *m* as follows:

$$AI_m = \sum_j R_j W_{mj} = \sum_j \frac{n_j}{\sum_m (P_m * W_{mj})} W_{mj}$$

where R_j represents the weighted hospital-to-population index of hospital *j,* n_j is the number of beds devoted to rehabilitation services of the hospital *j* and P_m is the resident population of the municipality *m*. W_{mj} that represents the weighting distance between the hospital *j* and the municipality *m* has been computed using the Sigmoid decay function. For each municipality *m* the percentage of *AI* related to hospitals located outside the belonging region was subsequently defined as:

$$AI_m^{extra} = \frac{\sum_{\{k\neq reg(m)\}} AI_m^k}{AI_m} = \frac{\sum_{\{k\neq reg(m)\}} \sum_{j\in\{Reg(j)=k\}} R_j W_{mj}}{AI_m}$$

where AI_m^k is the component of AI_m towards hospitals located in region *k*. For each region *i*, the AI_i^{extra} was computed considering the average value of AI_m^{extra} weighted by population.

In addition, a dispersion index (DI_i) that assesses the average distance travelled by patients from their region of residence is proposed as the weighted arithmetic mean of the accessibility indices AI_i^k weighted by passive mobility indices PM_i^k:

$$DI_i = \frac{\sum_{k\in\{k\neq i\}} AI_i^k * PM_i^k}{\sum_{k\in\{k\neq i\}} PM_i^k}$$

Data on hospitals was gathered from the Ministry of Health (MoH) website [13], while demographic data was collected from the Italian National Institute of Statistics (ISTAT) website [14]. All data refers to the year 2019, the most current information published by MoH. Travel distances were computed using the OSRM (Open-Source

Routing Machine) API [15]. Note that islands including Sardinia and Sicily were excluded from the analysis as residents cannot access to extra-regional facilities by car.

3. Results

The map shown in Figure 1 highlights the AI_i^{extra} computed for each municipality also reporting the total number of beds available in each municipality using black circles sized in proportion to the number of weighted hospital-to-population index (R_j). Moreover, passive mobility and accessibility indicators are reported in Table 1.

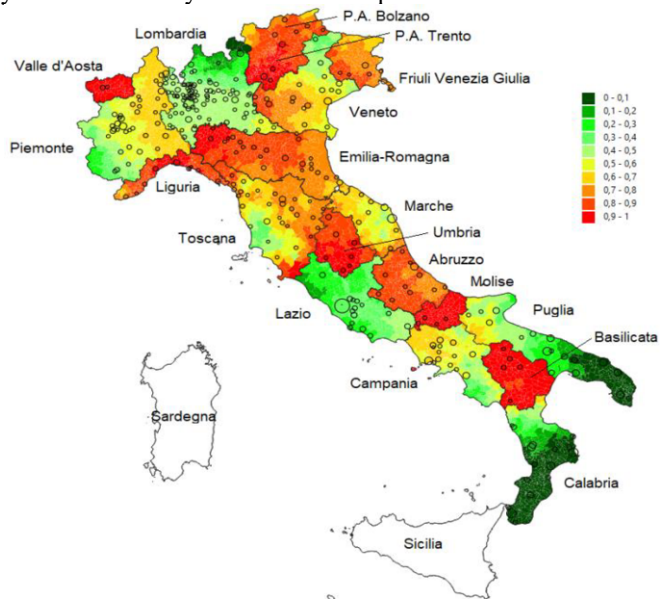


Figure 1. Extra-regional Accessibility Index (AI_i^{extra}) for rehabilitative beds in the Italian territory. Map defined and displayed using the open-source geographic information system (GIS) application QGIS [16].

To analyse how the different distributions of rehabilitative beds effects patient mobility, Figure 2 reports a scatterplot diagram that correlates it with the dispersion index. Firstly, all central-southern regions with the exception of Lazio are classified in the upper-left part of the diagram indicating a passive mobility higher than the national one and a mobility flow mainly towards health facilities which are distant from the region of residence. Basilicata and Molise, that also represent the regions with the greatest passive mobility (respectively 46.1% and 44.1%), are the two most remarkable examples, with over half of the total passive mobility directed to very distant hospitals (AI_i^k close to zero). This indicates that the proximity factor only partially influences the dynamics of passive mobility of patients residing in these regions. Note that the two regions with a lower dispersion index (Calabria and Puglia) have a predominant mobility (higher than 90%) to northern regions, such as Lombardia. The right part of the diagram comprises regions with passive mobility satisfied by health facilities located in neighbouring regions. This indicates that the proximity to structures located in non-resident regions plays a crucial role in determining patients' mobility. Among them Friuli Venezia Giulia reports the highest mobility rate (38.0%) almost entirely (34.0%) satisfied by facilities with wide extra-regional accessibility, such as Veneto and Lombardia. A similar pattern

is shown in regions with a low passive mobility rate (lower-right quadrant), such as Piemonte where 75% of mobility occurs in Lombardia with the rest provided by other close regions such as Liguria.

Table 1. Average values of Accessibility Index (AI_i) and its extra-regional component (AI_i^{extra}), the number of beds (PL_i), the passive mobility (PM_i) and the dispersion indices (DI_i). Data refers to the year 2019.

	Region	AI_i^{extra} (%)	PM_i (%)	DI_i	AI_i	PL_i
Central-southern regions	Abruzzo	82.0	21.5	10.73	27.71	44.10
	Basilicata	92.1	46.1	16.94	23.72	34.67
	Calabria	11.6	26.4	0.46	20.44	40.84
	Campania	64.5	24.8	6.88	37.23	31.05
	Lazio	36.0	11.0	3.02	39.35	48.29
	Marche	62.5	28.4	8.82	14.75	48.69
	Molise	95.3	44.1	9.58	25.06	45.24
	Puglia	23.6	20.9	0.39	21.93	29.89
	Toscana	71.7	20.9	11.80	20.12	53.48
	Umbria	89.8	22.7	12.03	28.72	47.46
Northern regions	Emilia Romagna	83.8	23.6	26.25	39.79	39.43
	Friuli Venezia Giulia	78.8	38.0	45.67	14.48	25.30
	Liguria	89.8	27.4	32.89	41.31	81.84
	Lombardia	46.4	9.3	13.72	63.35	34.99
	P.A. Bolzano	87.1	7.1	20.93	14.14	83.12
	P.A. Trento	92.4	15.4	25.36	29.32	88.56
	Piemonte	59.5	10.4	36.09	51.84	38.49
	Valle d'Aosta	98.2	20.7	41.60	41.34	74.76
	Veneto	69.4	14.7	18.97	35.38	46.00
	Italy	**57.2**	**15.9**	**18.2**	**38.60**	**62.95**

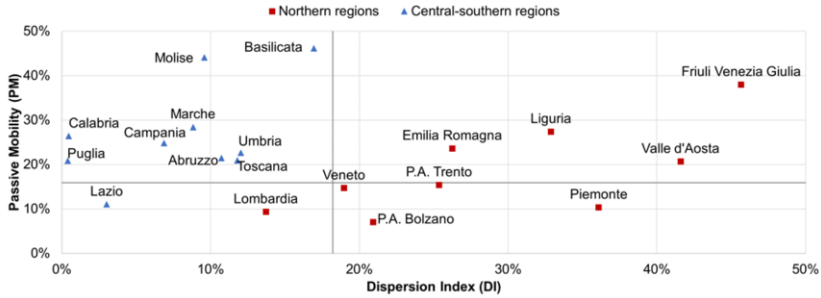


Figure 2. Scatterplot diagram reporting the Passive mobility index (PM) and dispersion index (DI)

4. Discussion

Different studies have analysed the main drivers of patient's mobility in Italy focusing on quality of services, social, demographic and economic status of patients and structural components. The aim of this study was to provide a preliminary analysis on the effect of hospital accessibility on patient's mobility across 19 Italian regions. This analysis allowed us to identify two main behavioural patterns. The first one entails regions whose patients accessed hospitals on the basis of their proximity. This is mainly found on northern regions where the request for hospital services may be associated with a capillary distribution network of extra-regional facilities that in regions, such as Friuli Venezia Giulia are easily accessible due to the conformation of the territory, the transport network and the distribution of population. The second pattern is composed by regions whose patients are willing or in need of travelling to access to hospitals located at very

long distances. This pattern is mainly found in southern regions, such as Molise and Calabria where a high percentage of patients access to extra-regional services provided by northern regions, such as Lombardia. This phenomenon can be associated to the demand of qualitative and timely services that are not always ensured by neighbouring facilities. However, this migration may be partly influenced by personal factors that are not strictly related with the quality of services, such as mobility for work or study reasons, the presence of family or relatives that can support patients during the hospitalization. In conclusion, this study confirms the main results reported in the literature highlighting that patients in the south of Italy when forced (i.e. long waiting times) or decide to migrate (i.e. searching for qualitative services) prefer to be hospitalized in facilities located in the north of the country than accessing those located in the neighbouring regions. While this finding highlights the potential inequalities both within and between regions, letting patients free to decide the professionals, facilities and places of care can be considered an opportunity to counterbalance socio-economic disparities, reducing unmet needs due to quality of territorial services and capacity of assistance. This preliminary analysis needs further investigation, for example, by analysing the mobility at meso and micro level and/or studying the effect of the extra-regional accessibility within a wider model that includes, for instance, socio-economic, demographic and other structural factors.

References

[1] Nante N, Guarducci G, Lorenzini C, Messina G, Carle F, Carbone S, Urbani A. Inter-Regional Hospital Patients' Mobility in Italy. Healthcare. 2021; 9(9): 1182.
[2] Fattore G, Petrarca G, Torbica A. Traveling for care: Inter-regional mobility for aortic valve substitution in Italy. Health Policy. 2014; 117(1): 90-97.
[3] Ricci A, Barzan E, Longo F. How to identify the drivers of patient inter-regional mobility in beveridgean systems? Critical review and assessment matrix for policy design & managerial interventions. Health Services Management Research. 2021; 34(4): 258-268.
[4] Berta P, Guerriero C, Levaggi R. Hospitals' strategic behaviours and patient mobility: Evidence from Italy. Socio-Economic Planning Sciences. 2021: 101030.
[5] Pecoraro F, Clemente F, Luzi D. The efficiency in the ordinary hospital bed management in Italy: An in-depth analysis of intensive care unit in the areas affected by COVID-19 before the outbreak. Plos one. 2020; 15(9): e0239249.
[6] Mafrolla E, D'Amico E. Patients' mobility as an indicator for (in)efficiency: a panel data analysis on Italian health care authorities. Health Econ Rev. 2013; 3:3.
[7] De Curtis M, Bortolan F, Diliberto D, Villani L. Pediatric interregional healthcare mobility in Italy. Italian Journal of Pediatrics. 2021; 47(1): 1-7.
[8] Balia S, Brau R, Marrocu E. Interregional patient mobility in a decentralized healthcare system. Regional Studies. 2018; 52(3): 388-402
[9] Pecoraro F, Luzi D, Clemente F. Spatial Inequity in Access to Intensive Care Unit Beds at Regional Level in Italy. Studies in Health Technology and Informatics. 2021; 281: 809-813.
[10] Pecoraro F, Luzi D, Clemente F. Analysis of the Different Approaches Adopted in the Italian Regions to Care for Patients Affected by COVID-19. Int J Environ Res Public Health. 2021; 18(3): 848.
[11] Ministry of Health. Report on hospital activity in Italy. 2019. Available from: https://www.salute.gov.it/portale/temi/p2_6.jsp?id=1237&area=ricoveriOspedalieri&menu=vuoto
[12] Aggarwal A, Lewis D, Mason M, Sullivan R, Van Der Meulen J. Patient mobility for elective secondary health care services in response to patient choice policies: a systematic review. Medical Care Research and Review. 2017; 74(4): 379-403.
[13] Ministry of Health. National healthcare service database. 2019. Available from: www.salute.gov.it/portale/documentazione/p6_2_8_1_1.jsp?lingua=italiano&id=6
[14] Italian National Institute of Statistics (ISTAT). Available from: https://www.istat.it/it/archivio/222527
[15] Open Source Routing Machine (OSRM). Available from: http://router.project-osrm.org
[16] QGIS. Available from: https://www.qgis.org/it/site

Challenges of Trustable AI and Added-Value on Health
B. Séroussi et al. (Eds.)

doi:10.3233/SHTI220557

Identifying Homogeneous Patient Clusters in Swiss University Hospital Through Latent Class Analysis

Gilles COHEN[a,1], Pascal BRIOT[b] and Pierre CHOPARD[b]
[a] *Finance Division, University Hospital of Geneva, Switzerland*
[b] *Quality and Patient Safety Division, University Hospital of Geneva, Switzerland*

Abstract. In hospitalized populations, there is significant heterogeneity in patients' characteristics, disease severity, and treatment responses, which translates into different related outcomes and costs. Identifying inpatient clusters with similar clinical profiles could lead to better quality and personalized care while improving clinical resources used. Super-utilizers (SUs) are one such a group, who contribute a substantial proportion of health care costs and utilize a disproportionately high share of health care resources. This study uses cost, utilization metrics and clinical information to segment the population of patients (N=32,759) admitted to the University Hospitals of Geneva per year in 2017 - 2019. Using Latent Class Analysis it identifies 8 subgroups with highly similar patients demographics, medical conditions, types of service and costs within groups and which are highly different between groups. As such 82% of all SU patients, 99% of all patients less than 20 years old and 78% of all orthopedics patients are clustered into only 3 separate groups while one group contain only adult women 90% of them 20 to 40 years of age.

Keywords. Latent Class Analysis, Clustering, Super-Utilizers, Inpatient Segmentation, Hospital Efficiency, Quality Improvement.

1. Introduction

The ongoing increase in healthcare expenditures [1] and the introduction of new payment incentives which favor reductions in avoidable admissions and reoperations [2] are forcing hospitals to develop new quality improvement strategies and improve their efficiencies. Since the greater share of hospital expenditures is often directed toward a limited number of patients commonly referred in the literature as super-utilizers (SUs) [3], identifying these patients as well as other homogeneous subgroups of patients and designing better targeted interventions for them have the potential to increase appropriateness of care, improve outcomes and reduce costs.

This study first expands upon prior work [4] to stratify the population of patients admitted per year from 2017 to 2019 to the University Hospitals of Geneva (HUG) applying a proven cluster analysis algorithm [4] on utilization data using demographics, admission and medical data. Second, it identifies and assesses the structural stability of

[1] Corresponding Author, Gilles COHEN; E-mail: gilles.cohen@hcuge.ch

inpatient subgroups at three different time points across the three years of hospitalization revealing similar clinical characteristics and patterns of care utilization.

The algorithm uses Latent Class Analysis (LCA) to identify patients' clusters within our inpatient data [4]. It is an iterative, maximum likelihood method that estimates how patterns in patient characteristics can be summarized into a finite number of groups, or latent classes, by producing a probability distribution over the cluster allocation for each patient. The present LCA analyses were not designed to track changes in inpatient class membership over time, but to examine whether certain patterns of hospitalized patients are consistent at different time points during hospitalization (i.e., structural stability).

2. Methods

2.1. Data and Variables

The HUG is the largest academic medical center in Switzerland with approximately 2,000 acute care beds and 47,000 admissions per year. All data for the study were collected from the HUG Enterprise Data Warehouse (EDW). The electronic medical record system provided clinical data while the accounting costing system provided financial data. Detailed admission data were gathered from hospital discharge summaries comprising admission and discharge dates, admission and discharge dispositions, lengths of stay (LOS), levels of care provided (standard care or intensive care), categories of services provided including surgical interventions, medications, tests, imaging and both primary and secondary diagnoses. The Elixhauser comorbidity index [5] was calculated for every admission using the International Classification of Diseases, Version 10 [6] using a coding algorithm. Diagnoses ICD-10 codes were matched with chapter headings and procedures codes were also reported. The top 10 percentile higher cost patients were identified based on their most expensive admission each year. For example, for the year 2019 we found that 90% of individuals (the least costly) accounted for only 58.5% of the total costs of the population while the remaining 10% of the population generated 41.5% of the total costs.

2.2. Methodology

From a technical point of view, LCA is a special kind of finite mixture model, also referred to as unsupervised learning models, which model a statistical distribution by a mixture (or a weighted sum) of several other distributions and clusters of similar data on the basis of selected parameters. The basic latent class model [7,8] postulates that there are underlying (latent) subgroups in a population that cannot be seen straight away, but must be inferred from a multiplicity of observations. Let y_j represent element j of a response pattern y. Let us establish an indicator function $I(y_j = r_j)$ that equals 1 when the response to variable $j = r_j$, and equals 0 otherwise. The probability of having a particular vector of responses is given as

$$P(Y = y) = \sum_{c=1}^{C} \gamma_c \prod_{j=1}^{J} \prod_{r_j=1}^{R_j} \rho_{j,r_j|c}^{I(y_j=r_j)} \tag{1}$$

where γ_c is the probability of membership in latent class c and is the probability of response r_j to item j, conditional on membership in latent class c. The γ parameters represent a vector of latent class membership probabilities that sum to 1. The ρ parameters represent a matrix of item-response probabilities conditional on latent class membership. Several criteria are available to determine from a set of models the best-fitting sub-class structure [9,10].

3. Results

A final LCA model of eight-class was identified based on a trade-off between several fitting indices, parsimony, and interpretability.

3.1. Segmentation outcome

Using the inputs variables described above in the clustering model 32,171, 32,450 and 33,157 unique patients across 8 groups were identified in 2017, 2018 and 2019 respectively by the clustering method. The number of patients per group ranges from 2,848 to 4,969 with an average of 4,074 across the study time period. For every year group 1 has only women patients and group 2 has only single patients as illustrated in figure 1. Over the study period admissions to the HUG were done more via the emergency department (ED) for all the groups (55.8%) with groups 6, 7 and 8 at 91.1%, 91.3% and 93.6% respectively. Group 2 was the exception with only 321 patients out of 12,011 (2.7%) admitted via the ED. These results are illustrated in figure 1. The patients' age showed a bimodal distribution with a first mode in the 0 to 18 age range (20.8%) and the second mode in the 75 and above age range (21.4%). Group 2 includes nearly only young patients less than 19 years of age (12,006 / 12,009). Group 7 has a majority (72.3%) of adults from age 75 and above.

Groups 1 and 5 show a range of precisely targeted procedures (such as obstetric technics and operations on musculoskeletal system) and primary diagnoses (such as pregnancy and delivery and diseases of the musculoskeletal system) illustrated in figure 1 while group 2 shows no procedures done in all three years. In addition, group 4 shows more than one third of patients with diagnoses of tumors.

The Elixhauser comorbidity index was calculated for each patient based on their diagnosis codes. The distributions per group for chronic heart failure, cardiovascular disease, chronic obstructive pulmonary disease, and diabetes do not show any significance difference across the groups.

Group 3 (N = 13,913) shows 63.8% of patients in the top 10 percentile for total costs. Group 3 patients includes the most number of patients with more than 10 ambulatory visits (42.6%), more than 10 different diagnoses (69.9%), more than 3 procedures (87.3%), more than 10 lab tests (74.5%), more than 10 medications (73.65), and more than 2 hospitalizations (45.3%). For each group the following financial measures were obtained given below for group 3 for illustration (see table 1). Note that the ratio of revenues and expenses is not an input variable of the clustering model.

Table 1. Financial measures

	2017	2018	2019
Patients count	4,553	4,766	4,594
Average costs	50,608	51,610	54,578
Average LOS	18.0	17.9	19.6
Revenues / expenses (R/E)	0.72	0.70	0.71
R/E when SUs patients excluded	0.92	0.89	0.85

4. Discussion and Conclusion

This study was conducted to determine whether a cluster grouping algorithm based on LCA clustering model could be applied to multiple years for the inpatient population of the HUG. The results show that the LCA clustering model is able to generate 8 groups with distinctive characteristics. In particular, the algorithm was able to identify a group with mostly patients less than 19 years of age who use the hospital for health related factors no hospitalizations for major conditions as well as a group with only women who are hospitalized for only women related procedures and diagnoses. A group of patients who most likely suffer from cancer related illnesses and a group of geriatric patients most likely suffering from multicomorbities and polypharmacy. Important among the 8 groups was the group of patients whose costs are in the top 10th percentile (Group 3) consistently for the 3 years and for whom the use of ambulatory and inpatient services is the greatest as well as the use of treatments, test (labs) and medications. Given the consistency of the results for these patients and the coherence we observed across the other groups (described above), we are confident that group 3 represents the SUs of care for the HUG in the years 2017 - 2019. Furthermore, the clustering method identifies 3 groups with highest yearly inpatient cost (including the SUs) whose revenues to expenses ratios were consistently in the 0.70 to 0.80 range for each year of study. Finally, across and among the groups the results for the variables studied are highly coherent demonstrating that the clustering algorithm is robust to stratify stable homogeneous populations for each year of the study. In this investigation, we extended the use of cluster analysis from our earlier work to identify distinct subgroups of patients across multiple years with specific combinations of co-occurring conditions in a large academic medical center. The model reveals highly stable expected segmentation by age brackets and gender such as with group 2 (patients less than 19 years of age) and group 1 (women only patients) along with the expected SUs care group 3. While the model appears coherent and robust to further assess the stability of these clusters over time, analyses should be conducted following a cohort of patients across multiple years. Such a study would provide additional characteristics of the groups we have identified such as for example when in time do patients move from one cluster to another, become SU of inpatient care, and what is their average yearly costs of care before and after they are identified as SUs.

In this study we showed how cluster analysis can be used to identify homogeneous groups of complex patients from a large heterogeneous population over a 3 years period. Such clustering method demonstrates that it is possible to use hospital inpatient data to systematically year after year identify homogenous grouping of patients with similar clinical and care utilization characteristics. Further studies in the work should also help us identify movements of patients across clusters across time. Recognizing and further studying these patterns of care usage would help hospitals design more personalized

patient centered care management strategies to better meet the needs of our patients while optimizing the use of our limited resources and controlling costs.

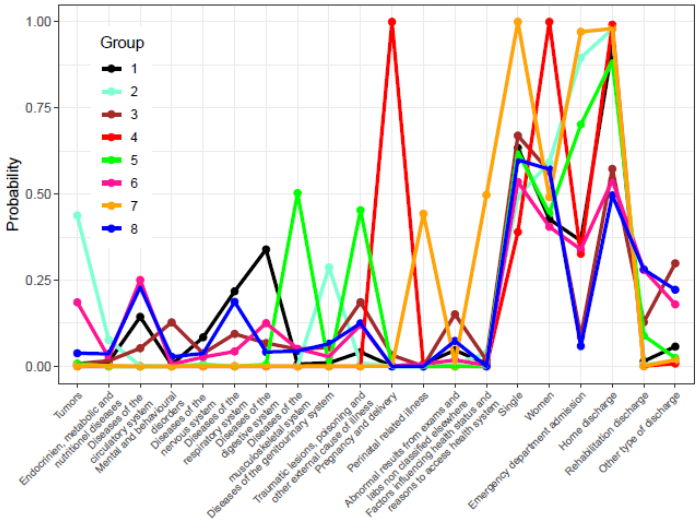


Figure 1. Profile plots for the 8-class model at time point 2019. On the x-axis are the diagnosis, demographics, admission and discharge characteristics variables with their response modalities. On the y-axis the conditional item probabilities are displayed, with each of the 8 latent classes being represented by a zigzagging line.

References

[1] Raghupathi V, Raghupathi W. Healthcare Expenditure and Economic Performance: Insights From the United States Data. Frontiers in Public Health. 2020;8.

[2] Purdey S, Huntley A. Predicting and preventing avoidable hospital admissions: A review. The journal of the Royal College of Physicians of Edinburgh. 2013 12;43:340-4.

[3] Wodchis WP, Austin PC, Henry DA. A 3-year study of high-cost users of health care. CMAJ. 2016;188(3):182-8.

[4] Cohen G, Briot P, Chopard P. Characteristics of Super-Utilizers of Care at the University Hospitals of Geneva using Latent Class Analysis. In: 9th International Conference on Computational Science and Engineering (CSE 2021). vol. 11; 2021. p. 600-16.

[5] Elixhauser A, Steiner C, Harris DR, Coffey RM. Comorbidity measures for use with administrative data. Medical care. 1998:8-27.

[6] Organization WH. The ICD-10 classification of mental and behavioural disorders : diagnostic criteria for research. World Health Organization; 1993.

[7] Clogg CC. Latent class models. In: Handbook of statistical modeling for the social and behavioral sciences. Springer; 1995. p. 311-59.

[8] McCutcheon AL. Latent class analysis. 64. Sage; 1987.

[9] Akaike H. A new look at the statistical model identification. IEEE Transactions on Automatic Control. 1974;19(6):716-23.

[10] Schwarz G. Estimating the Dimension of a Model. The Annals of Statistics. 1978 Mar;6(2):461-4.

Challenges of Trustable AI and Added-Value on Health
B. Séroussi et al. (Eds.)

doi:10.3233/SHTI220558

Using FHIR to Support COVID-19 Vaccine Safety Electronic Case Reports in America

Daniel A. RIZZATO LEDE[a,1], Helvert F. MOLINA[a], Maria Paz BERTOGLIA[a], Daniel OTZOY[d], Luis A. BENAVIDES[a,d], José A. DONIS[a], Víctor B. OSORIO[a], Carlos AGUILAR[a], José L. BUSTOS[a], Martín DIAZ[a], Karina REVIROL[e], César GALINDO[a,c], Jorge MANSILLA[a,c], Fernando CAMPOS[b], Diego KAMINKER[b], Marcelo D'AGOSTINO[a] and Desiree PASTOR[a]
[a] *PAHO,* [b] *HL7 Argentina,* [c] *HL7 Chile* [d] *RECAINSA,* [e] *MoH Argentina*

Abstract. During the COVID-19 pandemic, the Pan American Health Organization (PAHO) promoted several activities to strengthen the countries' emergency response. Vaccines represented a breakthrough in the pandemic evolution, even though they have not been equitably distributed. As most vaccines have received emergency authorizations for their timely delivery, vaccine safety surveillance has been highlighted for detecting early signals of potential adverse events following immunization (AEFI, also known as ESAVI). The objective of this article is to share the different steps, methodologies, and preliminary results of a regional policy to strengthen the ESAVI surveillance system in the Americas, including the adoption of HL7 FHIR for health information exchange between countries and PAHO.

Keywords. vaccine safety, adverse events following immunization, events supposedly attributable to vaccination or immunization, fast healthcare interoperability resources, health information exchange, interoperability.

1. Introduction

Vaccination is one of the most cost-effective methods for the prevention of infectious diseases [1]. Effective vaccines (i.e. vaccines inducing protective immunity) may produce some undesirable side effects which are mostly mild and clear up quickly. An Adverse Event Following Immunization (AEFI), also known as Event Supposedly Attributable to Vaccination or Immunization (ESAVI), is any untoward medical occurrence which follows immunization, and which does not necessarily have a causal relationship with the usage of the vaccine [2]. ESAVI surveillance is one of the most important activities to assure that vaccine products are safe and are being safely administered. Severe reactions following immunization are extremely rare, so several countries have joined forces to pool their ESAVI data in regional and/or global databases.

The COVID-19 pandemic promoted a race for the development of more than 100 vaccine candidates using a wide variety of platforms, some of them being newly available to the public, as the mRNA vaccines. The urgent need for a permanent prevention strategy in response to the pandemic required to fasten not only the development but the regulatory process. Data on efficacy, safety, and quality of manufacturing was limited during emergency authorizations and new strategies to ensure a vaccine-product risk-benefit continuous assessment should be in place. Moreover, considering the new challenges on safety and efficacy that innovation brought, steps were

[1] Corresponding Author, Daniel A. Rizzato Lede, Pan American Health Organization, Washington DC, USA; E-mail: rizzatodan@paho.org.

needed to develop robust surveillance systems that could effectively and efficiently identify immunization safety signals to facilitate public health decision-making. The process of detecting and reporting vaccine safety data has proven challenging, especially in the Americas.

The Uppsala Monitoring Center (UMC) is a World Health Organization (WHO) Collaborating Center that provides training, guidance, and support to countries in the WHO Programme for International Drug Monitoring [3]. UMC also manages VigiBase, WHO's database of individual case safety reports (ICSR) and the world's largest repository of adverse effects from medicines. Nevertheless, it was noted that only 13 countries from America shared vaccine safety information with UMC in 2021.

The Pan American Health Organization (PAHO) promoted several activities to strengthen the member state's emergency response to the pandemic. Regarding vaccine safety, the need for a regional approach was backed by the PAHO's Technical Advisory Group (TAG), WHO and the United States Center for Disease Control and Prevention (CDC). PAHO proposal included a regional vaccine safety surveillance network, based on the strengthening of national information systems to detect, notify, analyze, and classify the arising ESAVI cases. It would not challenge nor interfere with UMC-WHO data gathering. On the contrary, the enhancement of local capacities would improve data sharing with both PAHO and WHO.

The objective of this article is to share the different steps, methodologies, and preliminary results of a regional policy to strengthen the ESAVI surveillance in the Americas, including the adoption of HL7 FHIR (Health Level Seven - Fast Healthcare Interoperability Resources) for health information exchange between countries and the Pan American Health Organization.

2. Materials and methods

The Pan American Health Organization (PAHO) is the specialized international health agency for the Americas [4]. PAHO engages in technical cooperation with its member states to fight communicable and noncommunicable diseases and their causes, to strengthen health systems, and to respond to emergencies and disasters. To advance these goals, PAHO promotes technical cooperation between its 51 countries and works in partnership with ministries of health and other government agencies, civil society organizations, other international agencies, universities, social security agencies, community groups, and other partners.

The COVID-19 ESAVI project had different phases, starting with a charter, and advocacy activities to support it. After reviewing the proposal, WHO and CDC grants backed its execution. A regional survey was performed to assess each country's infrastructure and capacity building related to vaccine safety surveillance. Then, a phased approach was encouraged to promote progressive improvements in the ESAVI surveillance process and supporting technologies, departing from different baselines. FHIR was promoted for health information exchange between countries and the Pan American Health Organization, including investments in countries' capacity building, implementation guide creation, and the FHIR server deployment at a central level.

3. Results

According to the results of the "Regional Survey on the status of the information systems for the ESAVI surveillance" that was carried out in 2020 by PAHO, 3 different categories could be identified (see Table 1).

Table 1. Regional Survey on the status of the information systems for the ESAVI surveillance.

Category	% of countries	Characteristics
A	62%	based on paper, using spreadsheets to manually aggregate & tabulate data
B	21%	fragmented information system with different unconnected databases
C	17%	robust web based national information system

This heterogeneity motivated a pragmatic approach to quickly get information from countries. As a first step (Phase 1) PAHO asked each country to share a copy of their existing ESAVI database "as is", regardless of their level of development, through a secure File Transfer Protocol (FTP) server. A prior anonymization process was requested to exclude sensitive identification data. Each country's Ministry of Health (MoH) designated a delegate for manually sending the database on a weekly basis. As of January 2022, 16 countries have sent data to PAHO.

The national databases had enormous differences regarding their structure (what data they send, and how it was organized) and heterogeneity in quality (completeness, consistency, coding, among others). Feedback documents were shared with each country including these observations, to generate a continuous improvement process. Using the available information, a manual analysis was carried out to build indicators for the Pan American Advisory Committee on Vaccine Safety (PACVAS) and for country feedback. One of the most important challenges was (and still is) the standardization of the databases, including standard codes, and the mechanisms for sharing and aggregating all the information at the PAHO ESAVI Regional Database, minimizing the workload impact at the country level.

Phase 2 included completing the situation diagnosis and supporting the countries in achieving a robust national ESAVI system, including the 33 recommended core variables, and favoring interoperability between national institutions and automated reporting to PAHO's ESAVI regional database. Therefore, 3 fundamental lines of action were carried out:

1. The deepening of the country baseline assessment, and technical assistance for their national ESAVI information systems improvement.
2. The adequation of the open-source DHIS2 Tracker system, intended for Category A countries as a national ESAVI surveillance system.
3. The adoption of FHIR for the automated reporting of ESAVI cases from each country's information system to PAHO.

Within each country, an advocacy process was carried out, promoting good practices, international standards, inclusion of core variables for the ESAVI surveillance, processes reengineering, among others, which enacted joint efforts with national technical teams.

Category A countries needed a public health information system for ESAVI surveillance. PAHO analyzed different options and decided to promote an open-source system called DHIS2 (District Health Information System 2). DHIS2 is a public global good supported by the University of Oslo (UiO), being a software platform for

strengthening integrated health information management [5]. UiO developed a specific Metadata Package for AEFI surveillance within its Tracker app, as requested by WHO. PAHO requested further changes to DHIS2, including the adoption of semantic standards, the inclusion of extra core variables following its regional manual, among others. There have been countries that have advanced in the use of DHIS2 as a national ESAVI system, such as Ecuador, which will also adopt it for its nominal immunization registry. Other countries in the region such as Paraguay and Bolivia are taking steps in the same direction.

On the other hand, Category B and C countries needed an integration effort to connect their existing systems. This could be achieved internally through a custom extract, transform and load (ETL) process, and using an interoperability approach to send information to the regional database. In this context, we proposed to adopt FHIR, the newest open standard for health information exchange created by HL7 International. Several American countries such as the United States [6], Canada, Argentina [7], Brazil [8], Chile, and Colombia are already using FHIR for public health data exchange. The WHO itself promotes FHIR as a standard for structuring SMART guidelines, as digital certificates for COVID-19 vaccination, among other use cases [9].

The FHIR adoption project for ESAVI Surveillance was based on 4 pillars: Team building; Training; Creation of the ESAVI FHIR implementation guide; and Centralized installation of the FHIR server and client/country support.

The multidisciplinary team was built including people from different PAHO areas: Vaccine Safety, Immunization, Pharmacovigilance, Evidence and Intelligence for Action in Health and Information Technology Services. It was complemented by a team of HL7 FHIR experts from Argentina and Chile, and DHIS2 experts from the UiO. Internal courses were taken to level up the FHIR knowledge.

In the training axis, HL7 FHIR courses were contracted in Spanish, Portuguese, and English for technical representatives from member states' MoH. More than 100 people participated online in the FHIR Fundamentals courses taught by HL7 to date.

Creating a FHIR implementation guide (IG) for ESAVI notification involved an effort to standardize the variables included in the ESAVI Regional Manual, to generate computable specifications. In this way, the FHIR messages sent by the countries to PAHO would have a defined structure, facilitating their automated reporting. In turn, it was proposed to perform a mapping with the E2B XML standard (recommended for the exchange of security reports) to allow bidirectional conversion, and therefore communication with national and global systems that use that standard (Vigiflow, Vigibase, etc). The IG code system proposal was also challenging as many countries have heterogeneous ways of representing core variables (as medications, vaccines, medical background, or adverse events), from plain text to different coding standards as SNOMED, ICD-10, ICD-11, MedDRA, WHODrug, WHO ART, among others. Our approach was to promote coding standards and help countries to adopt them but allowing plain text descriptions in the meantime. There is a need of mapping those standards for coding transformation. Additionally, licensing of selected standards is an issue.

Finally, the installation of a centralized HL7 FHIR server is being carried out to receive and manage the information reported by the countries. At this stage, a group of pilot countries will adopt the FHIR API in their own ESAVI systems with PAHO's support. In this way, the Information System of each country (client) will securely authenticate and send a standardized FHIR message to the PAHO FHIR server, where the data will be received, stored, and processed.

4. Discussion and Conclusion

According to the information systems survey carried out by PAHO in 2020, 83% of the countries in the region DO NOT have information systems mature enough for ESAVI surveillance (categories A and B). Systems with fragmented data were found in multiple national institutions, and there was a lack of semantic standards for vaccines and associated adverse events.

Even though there is a previous standard for Individual Safety Case Reports (ICSR) called E2B XML, it has not been widely adopted for Vaccine Safety reports in the region. Some restrictions could be related to poor national information systems for pharmacovigilance, without semantic standards in place. Additionally, E2B only works for ICSR, requires licensed WHODrug and MedDRA codes, it is quite complex, and has a small community of practice. On the contrary, FHIR is a flexible, simple, easy to adopt standard that works for many health information exchange projects, not only pharmacovigilance. It has a big community of practice, and it is backed by several countries and public health organizations as the CDC and the WHO. Nevertheless, the question is not to replace but to add both standards for suitable circumstances.

Among the collateral benefits of the FHIR project, the capacities acquired by PAHO and by the countries in its management will facilitate its application to any other public health information exchange project. Many of the deliverables of this project can be reused, optimizing successive efforts. Each implementation guide will have its own scope, whether for individual reporting or aggregated data. This vision is aligned with the PAHO Roadmap for the Digital Transformation of the Health Sector in the Region of the Americas [10], and the WHO SMART guidelines proposal [11].

The project using FHIR for ESAVI electronic reporting is a proof of concept (PoC) to demonstrate the impact of this standard on regional public health data management. Future research and evaluation should add evidence on this subject.

References

[1] Bärnighausen T, Bloom DE, Cafiero-Fonseca ET, O'Brien JC. Valuing vaccination. Proceedings of the National Academy of Sciences. 2014; 111(34): 12313-12319. https://doi.org/10.1073/pnas.1400475111
[2] Adverse events following immunization (AEFI), (n.d.). https://www.who.int/teams/regulation-prequalification/regulation-and-safety/pharmacovigilance/health-professionals-info/aefi
[3] UMC, (n.d.). https://www.who-umc.org
[4] Who We Are, (n.d.). https://www.paho.org/en/who-we-are
[5] Koumamba AP, Bisvigou UJ, Ngoungou EB, Diallo G. Health information systems in developing countries: case of African countries. BMC Med. Inform. Decis. Mak. 2021; 21(1): 1-10.
[6] Helios, (n.d.). http://www.hl7.org/helios/
[7] Rizzato Lede DA, Pedernera FA, López E, Speranza CD, Guevel C, Maid JJ, Mac Culloch P, Rolandi F, Ayala F, Abadie DA, Baqué MI, Gassino F, Campos F, Kaminker D, Cejas CA, López Osornio A, Rubinstein A. Argentinian Digital Health Strategy, Stud. Health Technol. Inform. 2020; 270: 818–822.
[8] [No title], (n.d.). https://bvsms.saude.gov.br/bvs/publicacoes/strategy_health_digital_brazilian.pdf
[9] FHIR.WHO.DDCC-VS\Home - FHIR v4.0.1, (n.d.). https://worldhealthorganization.github.io/ddcc/
[10] CD59/6 - Roadmap for the Digital Transformation of the Health Sector in the Region of the Americas, (n.d.). https://www.paho.org/en/documents/cd596-roadmap-digital-transformation-health-sector-region-americas
[11] Mehl G, Tunçalp Ö, Ratanaprayul N, Tamrat T, Barreix M, Lowrance D, Bartolomeos K, Say L, Kostanjsek N, Jakob R, Grove J, Mariano Jr B, Swaminathan S. WHO SMART guidelines: optimising country-level use of guideline recommendations in the digital age. Lancet Digit Health. 2021; 3(4): e213–e216. https://doi.org/10.1016/S2589-7500(21)00038-8.

Challenges of Trustable AI and Added-Value on Health
B. Séroussi et al. (Eds.)

doi:10.3233/SHTI220559

The Tiel Shift Towards Person Centered Care

Jacob HOFDIJK [a, 1], Felix CILLESSEN [b] and Javier ASIN [c]
[a] *Casemix-CQT Zorg en Gezondheid, Utrecht, The Netherlands*
[b] *Hospital Rivierenland Tiel, the Netherlands*
[c] *ICSS, Paramaribo, Suriname*

Abstract. The health system in the Netherlands is one of the best in the world and it is a challenge to keep this affordable and accessible for everyone. A shift from care to lifelong maintenance of health is highly needed, but the drivers of change are missing. Obviously, the involvement of citizens is essential, but to introduce the change the indispensable incentives must be identified. The believe in integrated network care is growing, but it requires a fresh view on the integration of prevention and care, the involvement of the "old players" to create a new governance model, the supporting IT and adequate funding arrangements. In the Tiel region the implementation of the INCA model for integrated care in the primary care setting might make the difference for this journey.

Keywords. Care continuity, integrated health care systems, health maintenance, problem oriented medical record, interoperability

1. Introduction

To be sustainable the health care system needs to shift its focus from only caring for sick patients to the lifelong maintenance of the health of the population. The latter fits with the long-term strategy of the general hospitals in the Netherlands which intend to shift from a healthcare institution to an organization that focuses on self-reliance and maintenance of the health of the citizens. In the emerging health care network integrated care will require new coalitions of providers, professional coordination, sharing of data and person-centered overarching care plans [1].

The hospital as health focused organization can play a crucial role in this paradigm shift. In the Tiel region the INtegrated CAre (INCA) approach to manage patients with chronic conditions has been piloted [2]. The INCA model is based on the national program for care standards for chronic diseases as managed by the National Health Care Institute (ZIN). It defines the appropriate care for patients with chronic diseases like Diabetes, COPD and CVRM. The INCA model integrates the care standards into an integrated care program, assessing the health issues and leading to the choice of stepped care modules. It provides a health issue spiderweb presenting the actual status of each health and social care issue as basis for the shared decision making about the stepped care modules leading to an individual care plan.

[1] Corresponding Author, Jacob Hofdijk, Casemix-CQT Zorg en Gezondheid, Utrecht, The Netherlands; E-mail: jacob.hofdijk@casemix.nl

2. Methods

Seven years ago, a retrospective pilot was conducted with the INCA model in the primary care setting in Tiel (ECT). This study demonstrated the potential of the INCA spiderweb for the shared decision-making at patient level, and the contribution to the management and contracting of chronic care. Recently, the ECT decided to prepare the step towards the prospective use of the INCA model. From January 18th to February 4th 8 GP's, 7 Practice Nurses GP's and 6 ECT Support staff have been interviewed by Asin, ranging from 30 to 60 minutes, about the current potential of the INCA approach to their practice.

3. Results

There is full support from ECT and the insurer to shift to a prospective pilot of the INCA approach. This process will be aligned with the current steps of the providers in the region to develop a new governance structure for regional care delivery and its special focus on the digital transformation towards a regional IT ecosystem. A pilot will be started where the INCA approach will be tested prospective with the creation of an actual problem score for each patient based on actual data from different sources. The holistic approach will be introduced in the workflow, leading to a prospective plan for each patient for health, disease, and social care issues to be dealt with by different providers.

4. Discussion

Key to this approach is to create the individual care plan based on the national adopted standard for appropriate care to prevent chronic diseases and their complications. The main challenge is to find the incentives for the implementation of these quality standards across the silos of the health care system. It requires a paradigm shift of the focus from disease to health and supporting changes of the governance and funding system. The implementation requires a regional data infrastructure to connect both health and social care providers and the patient. Examples exists of the desired IT infrastructure [3].

5. Conclusion

The hospital is in parallel working closely with its regional partners to create a regional governance structure and digital transformation. So first small but important steps will be made for the transition to patient centered care not only leading to better health outcomes, but also better health equity for at least the population of the Tiel region.

References

[1] Cillessen FH, Hofdijk J. Transition Requirements from Problem List to an Overarching Care Plan for the Support of Person-Centered Care. Stud Health Technol Inform. 2020;272:292-5.
[2] van der Heide I, Snoeijs S, et al. Patient-centeredness of integrated care programs for people with multimorbidity. Results from the European ICARE4EU project. Health Policy. 2018;122(1):36-43.
[3] Serbanati LD, Ricci FL, Mercurio G, Vasilateanu A. Steps towards a digital health ecosystem. J Biomed Inform. 2011;44(4):621-36.

Challenges of Trustable AI and Added-Value on Health
B. Séroussi et al. (Eds.)

doi:10.3233/SHTI220560

Covid-19 Positivity Differences Among Patients of a Rural, Southern US State Hospital System Based on Population Density, Rural-Urban Classification, and Area Deprivation Index

Lori WONG, MD, MPH [a,1], Feliciano YU, MD [a],
Sudeepa BHATTACHARYYA, PhD [b], and Melody L. GREER, PhD [a]
[a] *University of Arkansas for Medical Sciences, Biomedical Informatics*
[b] *Arkansas State University, Biological Sciences Biosciences Institute, Jonesboro*

Abstract: In this study we examined the correlation of COVID-19 positivity with area deprivation index (ADI), social determinants of health (SDOH) factors based on a consumer and electronic medical record (EMR) data and population density in a patient population from a tertiary healthcare system in Arkansas. COVID-19 positivity was significantly associated with population density, age, race, and household size. Understanding health disparities and SDOH data can add value to health and the creation of trustable AI.

Keywords. Social Determinants of Health, COVID-19, Area Deprivation Index, Rural-Urban classification

1. Introduction

Arkansas' rural population face more socio-economic challenges than urban areas, which may increase COVID-19 infection risk. [1] The objective of this study was to assess the association of Area Deprivation Index (ADI), population density, and demographic factors on COVID-19 positivity among the urban versus rural subjects.

2. Methods

The study population consists of patients active in the tertiary healthcare system in 2020, defined as patients seen within the last 3 years (2017 to 2020) collected from the electronic medical record (EMR) in 2020. The study's SDOH (social determinants of health) data was from the Acxiom database which is composed of compiled consumer data. It was linked with EHR data collected in 2020 using name, address and date of birth.

[1] Wong, Lori, Department of Biomedical Informatics, University of Arkansas for Medical Sciences, 4301 W. Markham St., #782, Little Rock, AR 72205-7199; E-mail: ljwong@uams.edu.

A Spearman rank correlation was performed on COVID-19 positivity by zip code, ADI national rank, and population density (number of people per square mile). The subject population was separated into urban and rural groups, and COVID-19 positivity was defined as a patient that had COVID-19 at any time (yes/no). A logistic regression was used to estimate the impact of SDOH and population density on COVID-19 positivity for urban and rural groups. An $\alpha < 0.05$ was used to determine significance.

3. Results

There was a statistically significant relationship between COVID-19 positivity and population density by zip code in urban and rural groups with higher population density correlated with increased COVID-19 percent positivity (rho: rural 0.22, urban 0.39). ADI national rank was not statistically significant. The odds of COVID-19 positivity were lower in renters (Odds Ratio (OR): urban 0.85, rural 0.75), older age groups (30-64 (OR: urban 0.31, rural 0.46), 65+ (OR: urban 0.12, rural 0.21)), and a household size of two (OR: urban 0.76, rural 0.51). However, COVID-19 positivity was higher in the rural, Black population compared to the urban population (OR: urban 0.3, rural 2.0). In the urban population, no children present (OR: 0.76) had statistically significantly lower odds.

4. Discussion

This study found a significant association between population density and COVID-19 percent positivity for both urban and rural areas. This reflects findings of the early 2020 COVID-19 pandemic. [2] Some of the SDOH factors did not follow the expected pattern of increased COVID-19 positivity with worsening SDOH conditions. [3] This might mean that rural and urban areas have different risk factors for COVID-19 infection, which requires more exploration as these differences may impact AI models and improve explainability. The limitations to this study are generalizability due to the study population being from one hospital system.

5. Conclusion

Understanding health disparities and the data collected is key to producing not only added value to health, but trustable AI.

References

[1] Miller W, Wheeler E. Rural Profile of Arkansas, in: Social & Economic Trends Affecting Rural Arkansas, 2021.
[2] Wong DW, Li Y., Spreading of COVID-19: Density matters, PLoS One 15 (2020), e0242398.
[3] Kitchen C, Hatef E, Chang HY, Weiner JP, Kharrazi H. Assessing the association between area deprivation index on COVID-19 prevalence: a contrast between rural and urban US jurisdictions, AIMS Public Health 8 (2021), 519.

Challenges of Trustable AI and Added-Value on Health
B. Séroussi et al. (Eds.)

doi:10.3233/SHTI220561

Improving Shared Decision-Making Using Cognitive Effort-Optimization

Armin DARVISHZADEH SAYADI[a], Karim KESHAVJEE[1, a, b], Helen MONKMAN[b], Aziz GUERGACHI[c, d], Alessia PAGLIALONGA[e]

[a] *Institute of Health Policy, Management and Evaluation, University of Toronto, Toronto, Canada*

[b] *School of Health Information Science, University of Victoria, Victoria, Canada*

[c] *Ted Rogers School of Management, Ryerson University, Toronto, Canada*

[d] *Department of Mathematics and Statistics, York University, Toronto, Canada*

[e] *National Research Council of Italy (CNR), Institute of Electronics, Information Engineering and Telecommunications (IEIIT), Milan, Italy*

Abstract. Diabetes Prevention Programs (DPPs) can prevent or delay type 2 diabetes (T2D). However, the participation rates in DPPs have been limited. Many individuals at risk of developing diabetes have difficulties making healthy choices because of the cognitive effort required to understand the risks, the role of biomarkers, the consequences of inaction and the actions required to delay or avoid development of T2D. We report on the design and development of a prototype digital tool that decreases cognitive effort for people at risk of developing T2D using the effort-optimized intervention framework.

Keywords. Shared Decision-Making, Effort-Optimized Intervention Framework

1. Introduction

Diabetes prevention programs are proven to prevent or postpone T2D. However, participation rates in those programs have been consistently low, at about 19% [1]. For some individuals, the cognitive task of deciding to embark on a diabetes prevention program requires too much effort. The effort optimization intervention framework has been used successfully to help patients make decisions when faced with competing choices. In this project we designed and developed a prototype of a digital tool that reduces cognitive effort using an effort-optimized intervention framework [2]. We describe the process, the choices we faced and how we resolved them.

2. Design Methods

We explored the continuum of required activities and constraints in the user decision-making process and applied the following steps for planning an effort-optimized

[1] Corresponding Author, karim@infoclin.ca.

intervention:

- Decision activities should avoid frustration and self-efficacy reduction.
- Necessary effortful activities should lead to sustainable assets.
- Nurture salience to increase chances of desired behaviors in the face of competition.

To reach our goal, several techniques for planning an effort-optimized intervention were utilized to facilitate decision activities (see Table 1).

Table 1. The techniques that are used for planning an effort-optimized intervention during activities.

Activities	Techniques
1. Communicating Health Risks to individuals at risk	1) Provide visual aids with color-coded categories; 2) Provide a clear takeaway message for results; 3) Signal if differences are meaningful; 4) Provide thresholds for concern and action; 5) Individualize the frame of reference by allowing custom reference ranges; 6) Provide conversion tools along with risks and options; 7) Ease of using data [3].
2. User Interface Design for Data Collection	1) Use appropriate font and color-coding; 2) Use dropdown menus and autofill when appropriate; 3) Use plain language medical vocabulary instead of jargon; 4) Reduce the number of mandatory fields and open text; 5) Consider user interests and experiences [4].
3. Flexible Goals	Define specific, measurable, achievable, relevant, and timely goals
4. Communicating Potential Outcomes	Use of prescriptive analytics to determine blood sugar levels and body weight following participation in DPPs [5].

3. Conclusion

We have developed a functional prototype (https://doi.org/10.5281/zenodo.5893359) for a digital tool that helps individuals at risk make personalized decisions about lifestyle interventions. We attempted to identify and understand the underlying factors affecting individuals' decision-making processes and deployed the effort-optimized intervention framework to turn our ideations into a prototype. The next stage is to conduct a usability test to evaluate the prototype before testing the tool with the population at risk.

References

[1] Centers for Disease Control and Prevention. CDC - about the program - national diabetes prevention program - diabetes DDT [Internet]. Centers for Disease Control and Prevention. 2018 [cited 2021Dec26]. Available from: http://www.cdc.gov/diabetes/prevention/about.htm

[2] Baumel A, Muench FJ. Effort-optimized intervention model: Framework for building and analyzing digital interventions that require minimal effort for health-related gains [Internet]. Journal of Medical Internet Research. 2020 [cited 2021Dec27]. Available from: https://www.jmir.org/2021/3/e24905/.

[3] Witteman HO, Zikmund-Fisher BJ. Communicating laboratory results to patients and families [Internet]. De Gruyter. De Gruyter; 2019 [cited 2021Dec27]. Available from: https://www.degruyter.com/document/doi/10.1515/cclm-2018-0634/html

[4] Vickers AJ, Chen LY, Stetson PD. Interfaces for collecting data from patients: 10 golden rules. Journal of the American Medical Informatics Association. 2020;27(3):498–500.

[5] Perreault L, Pan Q, Mather KJ, Watson KE, Hamman RF, Kahn SE. Effect of regression from prediabetes to normal glucose regulation on long-term reduction in diabetes risk: Results from the diabetes prevention program outcomes study. The Lancet. 2012;379(9833):2243–51.

Challenges of Trustable AI and Added-Value on Health
B. Séroussi et al. (Eds.)

doi:10.3233/SHTI220562

Automated Twitter Extraction and Visual Analytics with Dashboards: Development and First Experimentations

Haris PATEL[a], Riyâz PATEL[a], Djamel ZITOUNI[b], Benjamin GUINHOUYA[a,b], Mathilde FRUCHART[a,b] and Antoine LAMER[a,b,1]
[a]*Univ. Lille, Faculté Ingénierie et Management de la Santé, F-59000, Lille, France*
[b]*Univ. Lille, CHU Lille, ULR 2694 - METRICS: Évaluation des Technologies de santé et des Pratiques médicales, F-59000 Lille, France*

Abstract. Information found in the social media may help to set up infoveillance and track epidemics, identify high-risk behaviours, or assess trends or feelings about a subject or event. We developed a dashboard to enable novice users to easily and autonomously extract and analyze data from Twitter. Eleven users tested the dashboard and considered the tool to be highly usable and useful. They were able to conduct the research they wanted and appreciated being able to use this tool without having to program.

Keywords. Data Science; Social media, Visual analytics; Education

1. Introduction

Information found in social media may help to set up infoveillance and track epidemics, identify high-risk behaviours, or assess trends or feelings about a subject or event [1,2]. In this study, we developed a dashboard to enable novice users to easily and autonomously extract and analyze data from Twitter. We assessed their ability to use it and captured their feedback on the value of such a tool for their daily practice.

2. Methods

The extraction of Twitter data is carried out using the open-source python package *snscrape*. The application is developed with *Dash*, an open-source framework for building web interactive applications entirely in Python. The dashboard allows the user to create a dataset, without coding, according to four criteria: the text query, the desired time range, the number of tweets to be retrieved and the language of the tweets.

The dashboard provides three tabs with (i) navigation across the dataset, and its exportation in CSV format, (ii) a time analysis with the count of tweets for each day and for each hour of the day, and (iii) a text analysis with a word

[1] Corresponding Author, Antoine Lamer, ULR 2694, 2 place de Verdun, F-59000, Lille, France; E-mail: antoine.lamer@univ-lille.fr

frequency graph and a word cloud [3]. We have proposed this tool to people who do not master programming (physicians, pharmacists and community managers), by instructing them to conduct a research with a keyword related to their fields of interest.

3. Results

Eleven users tested the dashboard (for topics such as supervised injection site, launching of a French prevention phone number about suicide, health policies, etc.) and considered the tool to be highly usable and useful. The figure 1 displays the wordcloud of frequency terms for query about supervised injection site. It highlights the towns in which supervised injection site are open or in the process of opening.

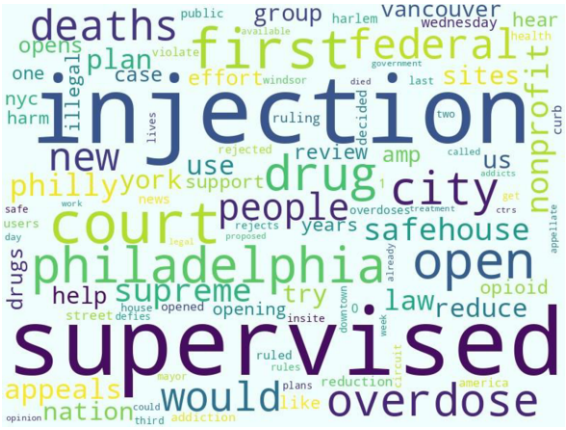


Figure 1. Wordcloud in the text analysis panel, for tweets about supervised injection site

4. Discussion and Conclusions

We offer a free tool which automatically extracts tweets and provides descriptive statistics. It can be used to follow the trends of current topics discussed on Twitter. With its search module, there is no need to program. The tool is available online and can be used for free.

References

[1] Abd-Alrazaq A, Alhuwail D, Househ M, Hamdi M, Shah Z. Top Concerns of Tweeters During the COVID-19 Pandemic: Infoveillance Study, J. Med. Internet Res. 22 (2020). doi:10.2196/19016.

[2] Castillo-Sánchez G, Marques G, Dorronzoro E, Rivera-Romero O, Franco-Martín M, De la Torre-Díez I. Suicide Risk Assessment Using Machine Learning and Social Networks: a Scoping Review, J. Med. Syst. 44 (2020) 205. doi:10.1007/s10916-020-01669-5.

[3] Twitter Analytics Dashboard, (n.d.). https://dash-ilis22.herokuapp.com/ (accessed January 26, 2022).

Challenges of Trustable AI and Added-Value on Health
B. Séroussi et al. (Eds.)

doi:10.3233/SHTI220563

Development Towards Patient-Centered eHealth Services in Finland

Marlon Luca MACHAL[1] and Alpo VÄRRI
Faculty of Medicine and Health Technology, Tampere universities, Finland

Abstract. The paper analyses the development of public eHealth services from 2014 to 2021 from the patients' point of view. The merits and missing features of the eHealth services were identified with patient interviews in 2014-2015. The list of missing features was again checked against the eHealth services in 2021. The main finding was that all the features wanted by the patients had still not been implemented. The finding of this paper suggests that current Finnish public eHealth services are organizations oriented rather than patient oriented.

Keywords. eHealth Services, Patient portal, Finland.

1. Introduction

In Finland, the eHealth services have been the responsibility of the over 300 municipalities in the country of 5,5 million inhabitants. Due to the large number of organisations providing these services, it is easy to notice the variation in the scope and quality of these services depending on the capabilities of the arranging organisation [1]. This study began in the year 2014 with the attempt to analyse the situation and to approach the eHealth services from the patients' point of view. The second part of the study in the year 2021 reviews how the eHealth services have developed, and do they now fulfil the patients' needs. This paper suggests the indication to continuously capture patients' needs.

2. Methods

The study began with an empiric survey of a number of websites for patients offered by the municipalities and hospital districts. Based on the observations of the websites study, an interview study was planned and performed in the years 2014 and 2015. This study surveyed the awareness of the people of the My Kanta [2] national eHealth service and presented its features to the respondents. Additionally, the respondents were given the opportunity to give feedback about the features of a patient portal based on a given example of a portal sketched by the researchers. In total, 351 people identifying themselves as patients were interviewed with a small number of questions. The 351 patients were informed about the purpose of collecting their feedback and that all information related to patients shall be confidential, their name, gender, age, exact location, and profession will be undisclosed. As result, patients contributed freely and

[1] Corresponding Author: Marlon Luca Machal; E-mail: marlon.machal@tuni.fi

showed high interest in this research topic. The interviews were conducted in the streets of a small number of cities in Finland. In 2021, a semi-structured interview was carried out virtually with three Finnish experts from Kanta who are responsible for My Kanta. The interview was used to verify that the collected voluntary feedback from 351 patients between the years 2014 and 2015 is still valid as of the year 2021. The views of Kanta experts are valuable because these experts receive continuous feedback from all users nationwide.

3. Results

The website survey gave the impression, that the offered eHealth services in 2014 were often designed with the top-down approach, i.e. what the organisation can provide instead of asking what the patients might want to have. The user interfaces varied significantly from one provider to another and a person moving to another location in the country would have to familiarize himself with a completely different user interface. Sometimes the websites had links that led to blank pages or errors. It was easy to see that improvements were possible. The voluntary feedback about eHealth services that was collected from the 351 patients was confirmed to be valid by the three experts of My Kanta service in June 2021. These three experts confirmed that they have received the same feedback from My Kanta users nationwide.

4. Discussion and Conclusion

Earlier studies [3, 4] have taken a top-down approach in assessing and suggesting solutions to improve Finnish eHealth services from health organizations' point of view. The outcome in the top-down approach aims to meet organizations' needs. This paper took a bottom-up approach where the design of eHealth services is based on patients' needs and the patients are active participants or drivers of the eHealth services. The major finding of this paper is that eHealth services offered by the Finnish public health care did not fully address the collected patients' feedback between the years 2014 and 2015. This study has shown that despite several years of development, the basic public eHealth services do not still fulfil all the central needs of the patients in Finland. This study underlines the importance of including the users of the services already in the specifications of included features of the eHealth service.

References

[1] EU-healthcare.fi (2021). Information about using health services in Finland and abroad. Available from: https://www.eu-healthcare.fi/healthcare-in-finland/healthcare-system-in-finland/

[2] Kanta (2021). My Kanta Pages. Available from: https://www.kanta.fi/en/my-kanta-pages (Accessed on 29.09.2021)

[3] Ruotanen R, Kangas M, Tuovinen T, Keränen N, Haverinen J, Reponen J. (2021). Finnish e-health services intended for citizens – national and regional development. Finnish Journal of EHealth and EWelfare, 13(3), 283–301. https://doi.org/10.23996/fjhw.109778

Jonsson F, Carson D, Goicolea I, Hurtig A. (2022). 'Strengthening Community Health Systems Through Novel eHealth Initiatives? Commencing a Realist Study of the Virtual Health Rooms in Rural Northern Sweden', International Journal of Health Policy and Management, 11(Special Issue on CHS-Connect), pp. 39-48. doi: 10.34172/ijhpm.2021.08

Challenges of Trustable AI and Added-Value on Health
B. Séroussi et al. (Eds.)

doi:10.3233/SHTI220564

Healthcare Insights: Evaluating the Access to the Italian Healthcare System

Monica BONETTO[a], Norbert MAGGI[a,b], Damiano FUSCHI[c], Alessandro VENTURI[d,e], Luisa BROGONZOLI[f] , Rosaria IARDINO[f], and Mauro GIACOMINI[b,1]
[a] *Healthropy s.r.l., Savona, Italy*
[b] *Dep. of Informatics, Bioengineering, Robotics and System Engineering, University of Genoa, Genoa, Italy*
[c] *Dep. of Italian and International Public Law, University of Milan, Milan, Italy*
[d] *Dep. of Political and Social Sciences, University of Pavia, Pavia, Italy*
[e] *I.R.C.C.S. Policlinico San Matteo, Pavia, Italy*
[f] *The Bridge Foundation, Milan, Italy*

Abstract. The Italian health system is organised on a regional basis and services are provided by both public and private operators, affecting the planning of services, access to services by citizens and their health rights. The creation of an observatory monitoring the methods and times of access to healthcare services has been pursued. The preliminary phase of the project is presented, which will lead to the comparison of the data obtained from 2019, with an eye on the Covid-19 pandemic impact.

Keywords. Public health, interoperability, waiting list data

1. Introduction

Monitoring of waiting lists has often been identified as an important factor in improving patient satisfaction [1-3]. Some studies suggest that waiting times may affect access to health services and that people with a higher socio-economic status wait less [4]. Services in Italian health system are provided by heterogeneous operators. This affects the choices by citizens, operators, political decision-makers, and service providers as relates to planning. Moreover, it affects equity of access to services by citizens and, therefore, their right to health itself. Healthcare Insight: Observatory of The Bridge Foundation (HI) focuses on monitoring access to the National Health System in a medium-long term perspective, creating free and public information.

2. Materials and Methods

This first phase of the project (started in February 2021) focuses on the collection and comparison of data on waiting lists for healthcare. Italian health organization focuses on the regions, but regions do not always make public data on waiting lists public.

[1] Corresponding Author, Mauro Giacomini, DIBRIS, University of Genoa, Via all'Opera Pia 13, 16145 Genoa, Italy; E-mail: mauro.giacomini@dibris.unige.it.

Bridge Foundation requested access to data (for 2019 and 2020) to the 21 Italian institutions through the Civic Access regulatory framework [5] that allows anyone to access public administration data, documents and information. The provided material is extremely heterogeneous. No entity has provided services to automatically obtain this data, although the Italian law states that this data must be provided in an open and interoperable way [6]. It was decided to organize this material in a relational data base in order to make data from different institution easily comparable and at all times.

3. Results

At the time of the presentation of this short article, the loading of the data into the DB has been completed, and an in-depth comparison is underway. During the upload, the collaboration of the individual entities in the project was evaluated taking into account accessibility, usability and completeness. Accessibility is characterised by the presence of a link for public access to the data and a built-in response to request. Usability is characterised by readability of the data by citizens and technical openness of the data. Completeness is characterised by data granularity, presence of data on outpatient services, presence of data of hospitalisation, presence of data on priority classes, frequency of data collections, data format.

4. Discussion and conclusion

Collaboration by the institutions in this first phase has been quite good (with significant peaks of excellence). The work that has been carried out so far, will be the basic for a more in- depth comparative examination of the data provided. Data comparing will also allow the impact evaluation of Covid-19 on the waiting lists.

5. Acknowledgments

HI is built thanks to the collaboration with the DIBRIS of the University of Genoa with the support of the Bank of Italy and UBI, the Italian Buddhist Union.

References

[1] Ko YK, The Relationships among Waiting Time, Patient's Satisfaction, and Revisiting Intention of Outpatients in General Hospital, Journal of Korean Academy of Nursing Administration, 2010 16.
[2] Xie Z and Or C, Associations Between Waiting Times, Service Times, and Patient Satisfaction in an Endocrinology Outpatient Department: A Time Study and Questionnaire Survey, INQUIRY: The Journal of Health Care Organization, Provision, and Financing, 2017 54.
[3] Netland P and McMullen, Wait time as a driver of overall patient satisfaction in an ophthalmology clinic, Clinical Ophthalmology, 2013.
[4] Simonsen NF, Oxholm AS, Kristensen SR, and Siciliani L, What explains differences in waiting times for health care across socioeconomic status?, Health Economics, 2020 29, 1764-1785.
[5] DECRETO LEGISLATIVO 14 marzo 2013, n. 33, in: art. 5, Italian Republic.
[6] DECRETO LEGISLATIVO 14 marzo 2013, n. 33, in: art. 7, Italian Republic.

Challenges of Trustable AI and Added-Value on Health
B. Séroussi et al. (Eds.)

doi:10.3233/SHTI220565

CovidGraph: Integrating COVID-19 Data

Ron HENKEL[a,1], Lea GUETEBIER[a] and Dagmar WALTEMATH[a]
[a] *Department of Medical Informatics, Institute for Community Medicine, University Medicine Greifswald*

Abstract. CovidGraph, developed by the HealthECCO community, is a platform designed to foster research and data exploration to fight COVID-19. It is built on a graph database and encompasses data sources from different biomedical data domains including publications, clinical trials, patents, case statistics, molecular data and systems biology models. The tool provides multiple interfaces for data exploration and thus serves as a single point of entry for data driven COVID-19 research.
Availability and Implementation: CovidGraph is available from the project website: https://healthecco.org/covidgraph/. The source code and documentation are provided on GitHub: https://github.com/covidgraph.

Keywords. COVID-19, Graph Database, Computational Biology Models, Ontology

1. Introduction

In 2020 SARS-COV-2 began to impact life across the globe on a scale unbeknown to humanity. Over the last two years fast and extensive research in that field generated a vast amount of knowledge about the virus. Research-wise, COVID-19 has been encountered with publications, patents, genome analysis, simulation studies for spread prediction, health studies and the extension of ontology information. One factor for fast and reliable research is commitment to the FAIR guiding principles [1].

CovidGraph offers **f**indable **a**ccessible **i**nteroperable and **r**e-usable COVID-19 data obtained, integrated and connected from open data resources. Data sets from the aforementioned domains are stored a graph database to offer researchers quick and efficient access to information about COVID-19. The connections within CovidGraph allow for new types of queries across previously disconnected aspects of the disease.

2. Methods and Results

Neo4J is a graph database engine designed to store and query a property graph in an efficient manner. It uses attributed and labelled nodes and relationships to represent the data. Textual information (e.g., publications, clinical studies or ontology terms) is analyzed using natural language processing and named entity recognition [2]. It is then semantically structured and represented as connected nodes. CovidGraph is based on a modular framework. For each data source, one ETL-process **e**xtracts the data, **t**ransforms

[1] Corresponding Author, Ron Henkel, Walther-Rathenau-Str. 48, Department of Medical Informatics, University Medicine Greifswald, D-17475 Germany; E-Mail: ron.henkel@uni-greifswald.de.

it using predefined mappings and loads the data into the CovidGraph. Subsequently, newly loaded data is cross-referenced and linked (by analyzing and matching publication and ontology terms) with data from already imported sources. CovidGraph comprises information about relevant publications from the COVID-19 Open Research Dataset [3], information about patents [4] and clinical trials [5]. Biomedical entities (e.g. genes, transcripts and proteins) are integrated from a variety of well-established databases [6]. Statistical data is imported from Johns Hopkins University [7]. Simulation models in standard format [8], including a Covid-19 model collection, are integrated from a domain-specific graph database (MaSyMoS [9]). The graph, as of today, contains 36 million nodes and 60 million relationships.

CovidGraph offers several interfaces for data exploration. The Visual Graph Explorer provides predefined views for an intuitive keyword-based graph exploration without prior knowledge of database query languages. SemSpect [10] supports drag & drop, expand and filter data items and automatic grouping of similar data items. Hence the graph can easily be traversed and visual representations can be created without detailed knowledge of the data model. Neo4j Bloom is an application for graph exploration. It offers semi-natural language queries, rule-based styling and search for phrases. The Covid-19 graph database can also be natively queried with pattern of interest. Query results are returned as a visual representation or as a tabular format or attribute-value pairs.

3. Discussion and Conclusion

CovidGraph is open source, FAIR and extendable. It provides a single-access point to previously disconnected data. CovidGraph integrates COVID-related data from the medical, biological and health domains into a knowledge graph. Using different interfaces, scientists can explore biomedical data domains in an easy fashion.

References

[1] Wilkinson, M, Dumontier, M, Aalbersberg, I, Appleton, G, Axton, M, Baak, A, Blomberg, N, Boiten, JW, Silva Santos, L, Bourne, P, others. "The FAIR Guiding Principles for scientific data management and stewardship". Scientific data 2016; 3(1):1–9.

[2] Lee, J, Yoon, W, Kim, S, Kim, D, Kim, S, So, C, Kang, J. "BioBERT: a pre-trained biomedical language representation model for biomedical text mining". Bioinformatics 2020; 36(4):1234–1240.

[3] Wang, L, Lo, K, Chandrasekhar, Y, Reas, R, Yang, J, Eide, D, Funk, K, Kinney, R, Liu, Z, Merrill, W, others. "CORD-19: the Covid-19 Open Research Dataset". arXiv 2020.

[4] The Lens. About The Lens. https://about.lens.org/covid-19.

[5] Zarin, D, Tse, T, Williams, R, Califf, R, Ide, N. "The ClinicalTrials.gov results database - update and key issues". New England Journal of Medicine 2011; 364(9):852–860.

[6] Henkel, R. "Biomedical Repositories for Simulation Studies". In: Reference Module in Biomedical Sciences. Elesvir, 2020. Available from: https://doi.org/10.1016/b978-0-12-801238-3.11684-8

[7] Dong, E, Du, H, Gardner, L. "An interactive web-based dashboard to track COVID-19 in real time". The Lancet Infectious Diseases 2020; 20(5):533–534.

[8] Malik-Sheriff, R, Glont, M, Nguyen, T, Tiwari, K, Roberts, M, Xavier, A, Vu, M, Men, J, Maire, M, Kananathan, S, others. "BioModels - 15 years of sharing computational models in life science". Nucleic Acids Research 2020; 48(D1):D407–D415.

[9] Henkel, R, Wolkenhauer, O, Waltemath, D. "Combining computational models, semantic annotations and simulation experiments in a graph database". Database 2015; 2015:1–6.

[10] Liebig, T, Vialard, V, Opitz, MConnecting the Dots in Million-Nodes Knowledge Graphs with SemSpect. In International Semantic Web Conference (Posters, Demos & Industry Tracks) 2017

Challenges of Trustable AI and Added-Value on Health
B. Séroussi et al. (Eds.)

doi:10.3233/SHTI220566

ILEG – A Case of Recording Patient Journeys

Julia S. VOLMERG [a,1] Jonas BIENZEISLER [a] Andrea D. KLAUSEN [a,b,c]
Insa SEEGER[b,c] Ulf GÜNTHER[b] Dominik DE SORDI [c] and Rainer RÖHRIG [a]
[a] *Institute of Medical Informatics, Medical Faculty of RWTH Aachen University, Aachen, Germany*
[b] *Research Network Emergency and Intensive Care Medicine, Oldenburg, Germany*
[c] *Carl von Ossietzky University, Oldenburg, Germany*

Abstract. In 2019 the Gemeindenotfallsanitäter was introduced in the area of Oldenburg, and scientific monitoring starting 2021 with Inanspruchnahme, Leistungen und Effekte des Gemeindenotfallsanitäters. Since then, it is possible to track patient journeys, starting from the emergency call to the subsequent treatment. This short communication provides an overview of the necessary data-acquisition and dataflow from all participating institutions and its possibilities.

Keywords. Emergency Service, Patient journey, Case Management, Medical Record Linkage, Community Health Workers, Collected Health Data

1. Introduction

In recent years, there has been an increase in ambulance calls and patients in Emergency departments (ED). Implementing community paramedicine has proven successful in relieving EDs worldwide [1]. The community paramedic (GNFS) was introduced in Oldenburg in order to relief these services and still provide adequate patient care [2]. In 2021 the ILEG study started to evaluate the GNFS-project. The objective of this study is to evaluate the quality of decisions, timelines, and efforts of the patient journey, from emergency call until the concluding treatment. This requires data collection on various points of care and merging them. This work focuses on the evaluation aspect patient journey within the ILEG-project.

2. Methods

The mapping of patient-journeys is made possible by linking data from different sources using a trusted third party (TTP). After the patient consents to participate, the documentation by the GNFS and the patient is registered and is asked to fill out a survey. Further sources are the family practitioner, the public-safety answering point (PSAP), the AKTIN Emergency Department Data Registry [3] and a telemedicine center. The registration is done with E-PIX from the MOSAIC-project [4]. Further on, the data is collected and assembled to represent the correct timeline, from initial emergency call, a possible visit to the ED, to the visit of the practitioner.

[1] Corresponding Author, Julia S. Volmerg, IMI RWTH UK Aachen, Pauwelsstraße 30, 52074 Aachen, Germany, E-Mail: jvolmerg@ukaachen.de

3. Results

Patient-journeys and the ILEG dataflow are shown in Figure 1. Until today 221 of 5474 (4%) patient-journeys were completely documented in 2021 and are currently analyzed in detail. The initial findings from the PSAP data, showing decreasing numbers in transports were similar to the findings from Tyano et al. [5], leading to lesser patients with non-emergency visits to the ED.

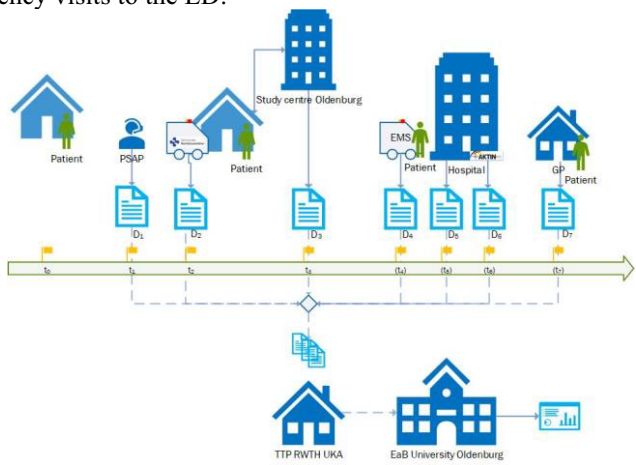


Figure 1: Timeline for a patient participating ILEG Project

4. Discussion and Conclusion

We were able to retrieve data from each participating institution and connect them by the identifying data to figure out the patient-journey. A limiting factor was the unexpectedly high number of patients unable to give a consent to the study. Usage data is limited for comparison across years due to the SARS-CoV II pandemic. Lessons learned for future is to lower access to participate and provide access for all relevant groups.

References

[1] Gregg A, Tutek J, Leatherwood MD, Crawford W, Friend R, Crowther M, McKinney R. Systematic Review of Community Paramedicine and EMS Mobile Integrated Health Care Interventions in the United States. Popul Health Manag. 2019;22(3):213–22. doi:10.1089/pop.2018.0114 Cited in: PubMed; PMID 30614761.

[2] Flake F, Schmitt L, Oltmanns W, Peter M, Thate S, Scheinichen F, Peters O. Das Konzept Gemeinde-notfallsanitäter/in. Notfall Rettungsmed. 2018;21(5):395–401. doi:10.1007/s10049-018-0426-7

[3] Ahlbrandt J, Brammen D, W. Majeed R, Lefering R, Semler SC, Thun S, Walcher F, Röhrig R. Balancing the need for big data and patient data privacy – an IT infrastructure for a decentralized emergency care research database. Studies in Health Technology and Informatics. 2014; Volume 205750–4. doi:10.3233/978-1-61499-432-9-750

[4] Bialke M, Bahls T, Havemann C, Piegsa J, Weitmann K, Wegner T, Hoffmann W. MOSAIC--A Modular Approach to Data Management in Epidemiological Studies. Methods Inf Med. 2015;54(4):364–71. doi:10.3414/ME14-01-0133 Cited in: PubMed; PMID 26196494.

[5] Tyano E, Ferrer E, Mayberry SD, Eldridge RC, Evans D, Simpson RL. Grady Health System's Mobile Integrated Health Program: A Statistical Analysis of Low-Acuity 911 Calls. JEMS Exclus. 2021;2021. Cited in: PubMed; PMID 34471915

Challenges of Trustable AI and Added-Value on Health
B. Séroussi et al. (Eds.)

doi:10.3233/SHTI220567

Identifying Determinants of Disparities in Lung Cancer Survival Rates from Electronic Health Record Data

Wanting CUI[1] and Joseph FINKELSTEIN
Icahn School of Medicine at Mount Sinai, New York, NY, USA

Abstract. The goal of this pilot study was to identify significant factors that affect disparities in lung cancer survival. A de-identified dataset was generated by querying electronic health records (EHR) from an academic medical center in New York City between January 2003 and November 2020. Socio-demographic characteristics, cancer stage, and genetic profile were analyzed using logistic regression. Two subsets of adult patients were identified: patients who were deceased less than 1 year after diagnosis and patients who survived over 5 years after diagnosis. Male, Black and Hispanic patients and those who were diagnosed in later stages were the people most susceptible to a shorter length of survival after cancer diagnoses. In addition, we identified three genetic oncodrivers (KRAS, EGFR and TP53) which were highly correlated with the length of survival after lung cancer diagnoses and their distribution was associated with race. We concluded that EHR data provide important insights on cancer survival disparities.

Keywords. Lung cancer, disparity, electronic health records, gene

1. Introduction

Lung cancer is one of the most common cancer in the United States for both males and females. It has a low five-year survival rate, compared to other common cancer types. Incidence and mortality rates are higher among Blacks as compared to Whites with lung cancer [1]. The major drivers of disparity in lung cancer survival include screening adherence, access to care and hereditary factors [2]. The goal of this pilot study was to identify major drivers that affect disparities in lung cancer survival using data from electronic health records (EHR).

2. Methods

A de-identified dataset was generated from Epic EHR system at the Mount Sinai Health System in New York City. We identified all adult lung cancer patients who were diagnosed between January 2003 and November 2020 within this dataset. We further identified 2 subsets of patients: patients who were deceased less than 1 year after diagnosis and patients who survived over 5 years after diagnosis. Logistic regression was performed to investigate the effect of demographic and cancer factors on patients'

[1] Corresponding Author, Wanting Cui, E-mail: wanting.cui@mssm.edu

duration of survival after cancer diagnosis. The independent variables comprised age, sex, race, cancer stage and genetic testing. The dependent variable was defined as a short-term survival. Somatic mutation data were used to identify genetic variants that affect lung cancer survival by analyzing variant distribution stratified by race and stage. We identified patients who had performed genetic testing and had test results in EHR.

3. Results

The analytical dataset contained 1099 patients: 750 of them survived over 5 years and 349 patients deceased within a year. In logistic regression, gender, race and cancer stage were important factors. However, age group and whether a patient has done genetic testing were not significant factors. Patients who were diagnosed at early stages of lung cancer were more likely to survive over 5 years than those who were diagnosed at advanced stages of lung cancer. In addition, Black patients and male patients had higher odds of being deceased in a shorter period of time after diagnoses, despite adjusting for age and cancer stage factors, compared to their White and female counterparts.

In genetic testing data, 214 (19.47%) patients had somatic genetic testing results. KRAS was the most common mutation. Around 47% of patients who underwent genetic testing had the KRAS gene mutation. In addition, 40% of patients who survived longer time had EGFR gene mutation, comparing to 12% of patients who survived a shorter time. Furthermore, patients who survived a shorter time have a higher proportion of TP53 gene mutations. There was higher proportion of Black patients (31%) than White patients (15%) with TP53 gene mutations who survived a shorter time.

4. Discussion

Male, Black and Hispanic patients who were diagnosed in later cancer stages were the people most susceptible to shorter length of survival after cancer diagnosis. KRAS was the most common genetic mutation among lung cancer patients and patients with TP53 mutations were at higher odds of being deceased in less than a year after cancer diagnoses.

5. Conclusions

Gender, race, cancer stage and somatic mutations were important factors that affects the length of survival of lung cancer patients after diagnoses. In addition, we found that Black patients had a higher proportion of the TP53 mutations, which was a gene mutation associated with short-term survival. Thus, future studies accounting for somatic mutations and availability of targeted treatment are warranted.

Reference

[1] Coughlin SS, Matthews-Juarez P, Juarez PD, Melton CE, King M. Opportunities to address lung cancer disparities among African Americans. Cancer Med. 2014 Dec;3(6):1467-76.
[2] Japuntich SJ, Krieger NH, Salvas AL, Carey MP. Racial Disparities in Lung Cancer Screening: An Exploratory Investigation. J Natl Med Assoc. 2018 Oct;110(5):424-427.

Challenges of Trustable AI and Added-Value on Health
B. Séroussi et al. (Eds.)

doi:10.3233/SHTI220568

Perception of the Communication Campaign for @choum a Symptom Reporting App: Insights from Semi Structured Interviews

Frederic EHRLER [a,1], Elias GAETE [b], David DE RIDDER [a b], Andrea LOIZEAU [a,b], and Idris GUESSOUS [a,b]
[a] *University Hospitals of Geneva, Switzerland*
[b] *University of Geneva, Switzerland*

Abstract. As an attempt to identify cluster of spread of COVID, we have developed the @choum functionality enabling individuals to signal when they perceived COVID-19 symptoms. The associated communication campaign did not encounter the expected success and only a limited amount of person did download the app. As an attempt to understand the barriers of use we have recruited a sample of general population to perform semi structured interview. Interview transcripts were analyzed using thematic analysis. Results highlight 3 profiles, engaged, critics and disengaged. We observe that these 3 profiles have different perception of the communication campaign, engaged participants being much more convinced by its message whereas disengaged people lack strongly of trust. This study helped us to identify what messages must be emphasized in order to attract critic people that may be convinced to use the tool.

Keywords. COVID, mobile app, epidemiology

1. Introduction

Digital technologies have played a central role to control the epidemic [1]. We have developed the @choum functionality embedded in the coronapp application, an app aiming at informing general population about COVID-19. It allows users to report when they notice symptom associated with COVID. All reports are processed by a cluster identification algorithm identifying clusters of possible Covid-19 outbreaks.

Several communication actions have been undertaken to promote the app. Among them flyers, targeted advertisements on social network and on the street as well as a promotional video. Beside the active communication, the app use statistics showed us a limited number of download and even more limited use of the functionality.

In order to understand the resistance factors among population to download and use the app we decided to perform interviews.

[1] Corresponding Author, Frederic EHRLER, University Hospitals of Geneva, Gabrielle-Perret-Gentil,4 1205 Geneve, Switzerland; E-mail: Frederic.Ehrler@hcuge.ch.

2. Method

The perception of the communication campaign by the target population has been captured using semi structured interviews including sociodemographic questions as well as questions about the perception of the COVID-19 crisis. It then presented the different communication actions undertaken to collect users' feedback.

Results were analyzed using a thematic analysis methodology to identify patterns of emerging themes [2]. Similar profiles were regrouped based on their perception on the situation and further thematic were analyzed in regards with these profiles

3. Result

Interview took place between 31 august and 15 September 2021. 10 participants were recruited. We analyzed the data and identified 79 initial codes for each relevant piece of information. Following this initial review, we looked for thematic among the codes by grouping the codes regarding their similarity into potential common themes (21 thematic). At this point, we condensed further the codes, guided by our research questions to draw out 3 general themes each containing sub-themes.

Based on the attitude and opinion of the participants, we did a simple cluster analysis to classify the participants under 3 profiles, engaged, critic and disengaged. The comparison of our 3 profiles regarding the effectiveness of the communication campaign showed clearly a much higher acceptance of the engaged group. Whereas the critic group did some constructing comments on the structure of the communication material, the disengaged group did not find any communication material convincing.

Discussion around the perceived messages of the communication campaigned highlighted the importance of being transparent regarding data protection. This is especially important since it is the major barrier that is reported by our sample that can take over the other benefits of the solution. Also all participants were unanimous to express that insufficient information were given on the

4. Conclusion

Although most engaged profile are receptive to our communication campaign, our message must be improved, especially in matter of trust, if we want to be able to convinced more critical persons. The intention to install the app is linked to the trust in the institutions and in the recommendations provided by the government. All these insight will help us to redesign our communication campaign and hopefully attract additional user to use the application.

References

[1] Vargo D, Zhu L, Benwell B, et al. Digital technology use during COVID-19 pandemic: A rapid review. *Hum Behav Emerg Technol* 2021; 3: 13–24.

[2] Kiger ME, Varpio L. Thematic analysis of qualitative data: AMEE Guide No. 131. *Med Teach* 2020; 42: 846–854.

Challenges of Trustable AI and Added-Value on Health
B. Séroussi et al. (Eds.)

doi:10.3233/SHTI220569

Symptom-Based COVID19 Screening Model Combined with Surveillance Information

Dohyung LEE [a], Myeongchan KIM [a], Hyunwoo CHOO [b], Soo-Yong SHIN [b, 1]
[a] *Mobile Doctor Co. Ltd, South Korea*
[b] *Department of Digital Health, SAIHST, Sungkyunkwan University*

Abstract. As the number of cases for COVID-19 continues to grow unprecedentedly, COVID-19 screening is becoming more important. In this study, we trained machine learning models from the Israel COVID-19 dataset and compared models that used surveillance indices of COVID-19 and those that did not. The AUC scores were 0.8478±0.0037 and 0.8062±0.005 with and without surveillance information, respectively, and there was significant improvement when the surveillance information was used.

Keywords. Machine Learning, COVID-19, Symptomatic screening

1. Introduction

As the number of confirmed cases of COVID-19 continues to grow unprecedentedly, effective triaging methods can mitigate the burden on healthcare systems. In previous studies, it has been reported that impressive detecting accuracy can be achieved even with a limited number of features through data publicly reported by the Israeli Ministry of Health [1,2]. However, epidemiological factors mainly influence the risk of being infected by contagious diseases. Thus, the symptom-based model has a drawback in that it does not consider epidemiological factors. In this study, we trained machine learning models with symptom data combined with surveillance information, which is one of the multiple epidemiological factors, and compared models which do not. Hence, our method can enhance the performance of machine learning models.

2. Methods

We utilized symptom data and PCR test results from COVID-19 Dataset by the Israeli Ministry of Health (Israel Dataset) and number of cumulative confirmed COVID-19 cases from Data Repository collected by the Center for Systems Science and Engineering at Johns Hopkins University (CSSE Dataset) from 2020.03.11 to 2021.12.31 [2,3]. Using the number of confirmed cases each day, we generated a national surveillance index and a global surveillance index by min-max normalization from the entire period for Israel and worldwide, respectively. We collated the dataset with 12 features in the following

[1] Soo-Yong Shin, 115, Irwon-ro, Gangnam-gu, Seoul, Republic of Korea; E-mail: sy.shin@skku.edu

list: sex, age (>=60 or <60 years), presence of 5 symptoms, (cough, fever, sore throat, shortness of breath and headache), binary values of 3 test indications ("contact with confirmed cases", "abroad" or "others"), and 2 surveillance indices on the PCR test date (national and global). Out of a total of 7,310,559 entries, we randomly sampled 300,000 entries and split 100,000 entries into training, validation and test datasets. We used a logistic regression model (LR) and 16 layers of 1-dimensional convolutional neural network model with skipped layer (1D ResNet). To see the effect of the surveillance indices on the models, we repeated experiments with and without the surveillance indices (12 features and 10 features, respectively) 5 times for each model. Then we performed a two-sample t-test to compare with and without the surveillance indices.

3. Results

The AUC score of logistic regression with surveillance indices was 0.8374±0.0019, and the AUC score was 0.7828±0.0031 without the indices. The score for 1D ResNet was 0.8478±0.0037 when the surveillance indices were included and 0.8062±0.005 when not included. In the presence of the surveillance indices, both models significantly improved the performance (P<.001). Figure 1. shows the average AUROC of both results.

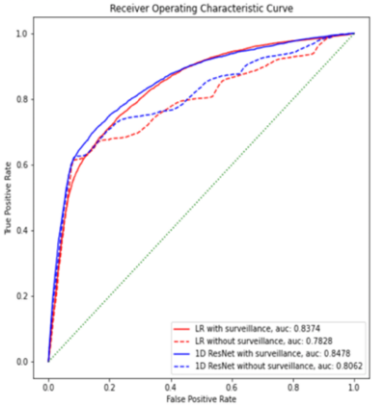


Figure 1.

4. Discussion and Conclusions

Performance of our two models for COVID-19 diagnosis prediction improved when the training data included national and global surveillance index. In the early epidemics, a prior study trained a model with 10 days of Israeli Data, but when this model was tested after more than a year, the performance was not as good written in the study [1]. This may be due to changes in epidemiological characteristics such as variant viruses, and vaccination rates. Since clinical assessment considers epidemiological factors, we also considered them in this study. In addition, we expect to improve performance by adding other epidemiological information such as vaccination rates and local surveillance information.

5. References

[1] Zoabi Y, Deri-Rozov S, Shomron N. Machine learning-based prediction of COVID-19 diagnosis based on symptoms. npj Digital Medicine, 2021;4(1),1–5.
[2] COVID-19-Government Data, 2020, https://data.gov.il/dataset/covid-19
[3] Dong E, Du H, Gardner L. An interactive web-based dashboard to track COVID-19 in real time. Lancet Inf Dis. 20(5):533-534

Challenges of Trustable AI and Added-Value on Health
B. Séroussi et al. (Eds.)

doi:10.3233/SHTI220570

CyberPharmacovigilance of Covid-19: Social Media Data Analytics

Eustache Muteba AYUMBA[a,1]
[a]*International Medical Informatics Association, Correspondent in DR Congo*

Abstract. In this article, we present a methodology and a tool for extracting and analyzing information, reported by a social media monitor, of people who have taken drugs to treat Covid-19 and the adverse effects encountered.

Keywords. Pharmacovigilance, Covid-19, Social media monitoring, Data analytics

1. Introduction

WHO's involvement in monitoring the situation of pharmaceutical industry is pushing the authorities in charge of the issue of organizing pharmacovigilance. Most of time, decisions in the area of drug safety are based on available information from pharmacovigilance systems, which are mostly based on Adverse Drug Reactions (ADR) reports, voluntarily made by healthcare professionals and consumers. In some cases, some decision-makers and health caregivers do not have up-to-date and well-organized pharmacovigilance systems. With the onset of the COVID-19 pandemic, social media has rapidly become a crucial communication tool for information generation, dissemination, and consumption. Thus, social media monitoring [1], along with conventional pharmacovigilance measures, can be used to detect signals associated with any information related to adverse effects or any other problem related to marketed medicines. This article presents a methodology and a tool for extracting and analyzing information, reported by a social media monitor, of people who have taken drugs to treat Covid-19 and the adverse effects encountered.

2. Methods

Our analyzer system is naturally based on existing works such as developed in [2, 3, 4 and 5]. It is composed by the following components: the social media monitor, the social media report file, the indexer, the search index, the index file and the dictionary, the user interface, the extractor engine, the pharmacovigilance database and the data analytics report . The Analyzer system tasks are the following. Build the indexer (index file) from the lexicons and set a dictionary; the user interface allows the user to submit queries; once the query is processed, the extractor engine sends the results to the user interface for validation; on needed refines information; stores it in the database and reports the data analytics result. The extraction and the analysis of data are based on a query text as

[1] Eustache Muteba Ayumba, P.O. Box 14769, 01, Kinshasa, Congo, E-mail: emuteba@hotmail.fr

decision criteria or rules. The extractor engine captures salient information and computes the term frequency (TF) of the indicators (keywords) retained in the analysis plan. We proposed the following query structure. *Query Structure (Expression form):*
(1) Categories: keywords and basic units that identify the conversation
(2) Lexicons: keywords that specify your query with semantic type avoiding noisy.
(3) Boolean terms: "AND, OR, and NOT" that create more powerful searches.

3. Results and Discussion

The social media monitoring tool used in our study is Alerti (https://fr.alerti.com/). Our analysis plan was based on query text containing keywords such as categories are "anti-covid", "side effect", "adverse drug reaction" and others keywords related to drug used as anti-covid. The social media monitor generates a report in .csv file. Our analyzer system, pharmavigil (https://www.vchf.net/pharmavigil/pharmav.php), was design and build to reinforce the data analysis, with the same query text, to get salient information. The experimentation was done on the social media monitor report (file). Our data analytics report generated by the extractor engine provides two kinds of information namely a list of record lines containing keywords and the term frequencies. We note, in case of anti-covid drug "ivermectin", the TF value is 1390 on a total 1756 of record lines. And a "nitazoxanide" drug has a TF value 0 on the same total records lines. The task of analyzing data and searching specific words into a social media monitoring report is subject to many challenges. One of the crucial challenges is a language style devoid of linguistic vigor. It should be noted that to overcome the limits imposed by the keyword search, our system uses the dictionary. This allows so little to eliminate the noises therefore the post messages containing irrelevant information.

4. Conclusion

On the strength of the above, we believe that in the field of health, social media can be a potentially exploitable source of information to complete and optimize pharmacovigilance processes. The main contribution of our study is the methodology of developing an analytics supplement tool that aggregate data coming from social media monitoring.

References

[1] Lee J, Lee Y, Kim DH, Lee HS, Yang BR, Kim MG. The Use of Social Media in Detecting Drug Safety–Related New Black Box Warnings, Labeling Changes, or Withdrawals: Scoping Review. JMIR Public Health Surveill. 2021 Jun;7(6):30137.
[2] Twetman H, Paramonova M, Hanley M. Social Media Monitoring: A Primer - Methods, tools, and applications for monitoring the social media space, NATO Strategic Communications Centre of Excellence. 2020.
[3] Podhoranyi, M. A comprehensive social media data processing and analytics architecture by using big data platforms: a case study of twitter flood-risk messages. Earth Sci Inform. 2021; 14:913–929.
[4] Mucchielli R. L'Analyse de Contenu - Des Documents et des Communications. 5ième ed. Paris : ESF; 1984.
[5] Levene M. An Introduction to Search Engines and Web Navigation. 2nd ed. New Jersey: John Wiley & Sons; 2010.

Section VI

Human Factors Organizational Issues; Societal Aspects and Education in Healthcare

This page intentionally left blank

Challenges of Trustable AI and Added-Value on Health
B. Séroussi et al. (Eds.)

doi:10.3233/SHTI220572

Personal Health Records an Approach to Answer: What Works for Whom in What Circumstances?

Elisavet ANDRIKOPOULOU[a] and Philip SCOTT[b]
[a] *University of Portsmouth*
[b] *University of Wales Trinity Saint David*

Abstract. National Health Service (NHS) policy suggests that increasing usage of electronic personal health records (PHR) by patients will result in cost savings and improved public health, especially for people with long-term conditions. PHR design features are inevitably important, since a good PHR design should make the users achieve their health goals effortlessly, which is understandable and usable. Three original theoretical models were developed using realist evaluation, one per long-term condition cohort, describing the interaction between the PHR design features and the patient and disease specific factors, to help determine what works for whom in what circumstances.

Keywords. Personal health record, realist evaluation, diabetes, asthma, heart disease

1. Introduction

Personal health records are defined as "online systems that include collections of patients' healthcare and medical data, which enable patients to share, organize and manage these data according to their own views" [1]. PHRs are typically important for patients that are suffering from chronic conditions, who gain the most value and have the higher adoption rate [2]. PHRs and other mobile health apps are popular, however only 15% of them are linked to some medical providers, who would potentially be able to implement human-computer-interaction (HCI) design and usability standards [3]. How can we ensure that a PHR meet the needs of their target audience? A good HCI-based PHR design should enable users to achieve their health goals effortlessly, whilst being usable [4].

Evaluating health technology is always a challenging and risky goal, especially when there is a clash of methods, frameworks and formal methods of the health technology assessments worldwide. This study is not researching how PHRs improve people's health but it is solely about patient reported factors that are able to make PHRs easily adopted and usable. The UK Medical Research Council guidance [5] recognises the need for evaluating complex interventions and cites realist evaluation as an important theory-based approach. There is no doubt that a PHR is a "complex" intervention: it comprises "multiple interacting components" and faces numerous barriers to adoption.

This study is part of a larger research project, the ePHRma study, where the most useful PHR design features to improve medication adherence in adults with long-term conditions are explored. Medication adherence can be defined as "the extent to which a

person's behaviour towards their medication intake, corresponds with agreed recommendations from a health care provider" [6]. Medication adherence is a well-known challenge in healthcare, and is related to a large number of factors such as side effects, forgetfulness or effective self-management and is affected by psychological factors and beliefs [7].

This paper aims to answer the following research question:

What theoretical model can describe and explain the interaction between the PHR design features and the patient and disease specific factors, to help determine what works for whom in what circumstances?

2. Methods

The methods are described in detail in the published protocol [8]. In this study, the qualitative data have been identified as dominant [9], as the area is largely complex; thus the sampling strategy was based on the qualitative component of the study. The qualitative strand used interviews to collect data from a purposive convenience sample of participants, which were analysed using the Framework method [10] to answer the research question. The 'gold standard' for a purposive sample is to achieve saturation, which is impossible to predict; thus, a typical recommendation is medium sample size, roughly 30–40 participants.

Realist evaluation was used to construct the theoretical models that answer the question "what works for whom, in what circumstances, what respects and how?" [11]. This study's realist evaluation begins by seeking an explanation for why the outcome of interest occurs for some people and not for others. This study identify mechanisms from a pragmatic scope, at the level of human reasoning [11]. We have defined the following categories of mechanism in health informatics: social, emotional state, economical, educational, environmental or cognitive. These umbrella terms were conceptualised using the COM-B model [12] and the Identification of Medication Adherence Barriers Model (IMAB) [13]. The study was given Research Ethics Committee (REC) favourable ethical opinion and final approval by the Health Research Authority (HRA) and Health and Care Research Wales (HCRW) on 10 September 2018. The REC reference is: 18/NE/0253, chaired by Professor Andrew Hall.

3. Results

The people with diabetes cohort has 17 participants. The majority of the participants have completed at least a bachelor's degree in higher education (n=14) and most of them are between 35-54 years old (n =8) but there are representatives from all relevant age groups. The heart disease cohort has 13 participants. There are more female (n=8) than male participants (n=5) and the majority of them (n=11) have high education status and high health literacy. The people with asthma cohort has 12 participants. The majority of the participants are younger than 54 years old (n =10) with only 2 participants aged 65 and over.

The final programme for the people with diabetes cohort is illustrated in Figure 1. Patients who consider their medication regimen too complex or not sufficient or feel less capable of handling their medication, blood glucose testing or planned activities or feeling unrest and having low confidence levels are less likely to adhere. Patients who

have better patient-clinician relationship and prescribed access to diabetic kits such as Dexcom or Freestyle Libre perceive their medication regimen as simple and are less resistant to change that leads to better medication adherence. Patients with a greater sense of goal commitment perceive their medication regimen as something that they have to do to maintain their health-related quality of life and are less resistant to change. The sense of control that a prescribed diabetic kit such as Dexcom or Freestyle Libre provides them increases their feeling of esteem. When an adult with diabetes is using a PHR in order to better plan their day or activities or injections and therefore, to improve their medication adherence and their health-related quality of life.

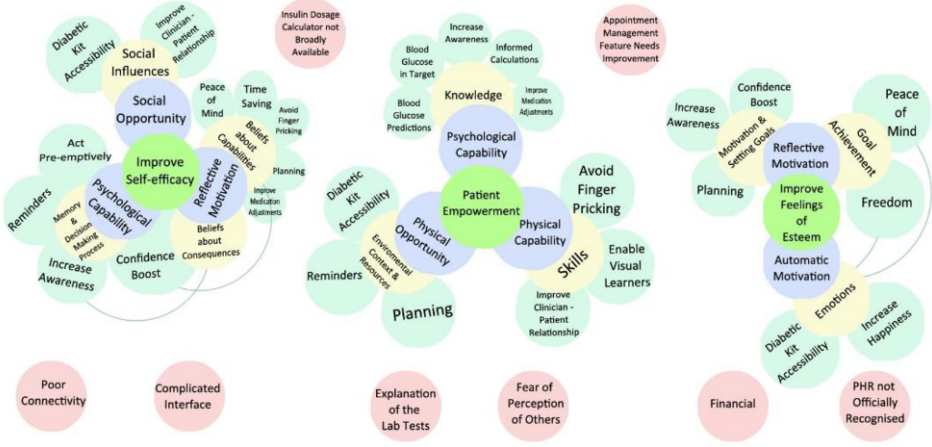


Figure 1. Final CMO configuration for people with diabetes cohort. Middle green: Mechanism; blue-purple: COM-B factors; yellow-orange: IMAB domains; light green: themes; red: barriers. Illustration from: StellaKAngel

Figure 2. Final CMO configuration for people with heart disease cohort. Middle green: Mechanism; blue-purple: COM-B factors; yellow-orange: IMAB domains; light green: themes; red: barriers. Illustration from: StellaKAngel

The final iteration of the programme for the people with heart disease cohort is illustrated in Figure 2. Patients may improve their current medication adherence, by viewing a medication adherence percentage and sometimes sharing it in their social media of choice. An easy to use, often wearable and convenient PHR may enable patients to better plan their lives thus to improve their medication adherence. Some participants of this cohort stated that using a PHR increases patient-clinician communication and trust, since data that were able to be shared between the patient and clinician provided the later with a more holistic knowledge of the effectiveness of the therapy the patient was following (or not following).

The final iteration of the programme for the people with asthma cohort is illustrated in Figure 3. For this cohort acting pre-emptively to contact a clinician or take their emergency medication seem more important than actually manage their medication dosages. They are confident that they are taking their medication correctly and their attitude portrays this, since are monitoring their symptoms, progression and reactions to change the social context and society's perception by being pleasers to their healthcare providers. By improving their confidence and lowering their stress levels patients are more focused on their health-related goals improving their medication adherence. Similarly, by being more active they may change their behaviours to improve their exercise regime and therefore their medication adherence, since the need for the use of emergency inhalers is lower.

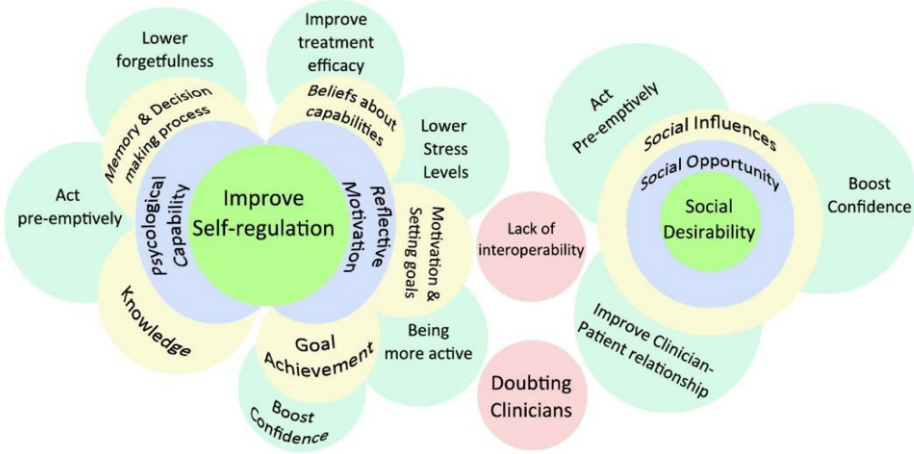


Figure 3. Final CMO configuration for people with asthma cohort. Middle green: Mechanism; blue-purple: COM-B factors; yellow-orange: IMAB domains; light green: themes; red: barriers. Illustration from: StellaKAngel

4. Discussion and Conclusion

This study used the realist evaluation and created three theoretical models. The diabetes and heart disease cohorts have the most similarities in terms of what mechanisms enable them to improve medication adherence and health related quality of life. The asthma cohort has different mechanisms for the same outcome. Interestingly the IMAB domains are intertwined but based on the interview analysis and themes, the mechanisms that are activated are different. This could be explained due to the different long-term conditions

and differences on their care. The difference could also lie to the way the patients are administering their medications, as people with heart disease and diabetes are able to measure a value related to their condition to assess whether they are adhering to their medication or not. For people with asthma this is not possible, since there is no easy way to measure a flair-up. Often people with asthma consider that their shortness of breath is "normal".

Participant recruitment was challenging in this study and due to the voluntary nature of participation, selection bias could occur. However, we have worked closely with a PPI group to develop the recruitment strategy ensuring that the study's recruitment targets and methods are realistic. Overall, the participants mentioned a number of improvements in medication adherence, compliance and overall attitude, whilst using a PHR. In relation with the "real world", this study showed that PHR developers should work per cohort and contextualise their applications based at least on a long-term condition.

References

1. Andrikopoulou E, Scott PJ, Herrera H. Important Design Features of Personal Health Records to Improve Medication Adherence for Patients with Long-Term Conditions: Protocol for a Systematic Literature Review. JMIR Res Protoc [Internet]. 2018;7(6):e159. Available from: http://www.researchprotocols.org/2018/6/e159/
2. Chrischilles EA, Hourcade JP, Doucette W, Eichmann D, Gryzlak B, Lorentzen R, et al. Personal health records: a randomized trial of effects on elder medication safety. J Am Med Inform Assoc [Internet]. 2014 [cited 2017 May 8];21(4):679–86. Available from: http://www.ncbi.nlm.nih.gov/pubmed/24326536
3. Yang S, Boulos MNK. Mobile Health (mHealth) Apps for Improved Health in Medicine [Internet]. Medicine 2.0'14 Summit & World Congress, At Sheraton Maui Resort, Maui, Hawaii, USA. 2014. Available from: https://www.researchgate.net/publication/268224586_Mobile_Health_mHealth_Apps_for_Improved_Health_in_Medicine
4. Norman DA. The design of everyday things. Revised an. Basic Books, a member of the Perseus Books Group; 2013.
5. Moore GF, Audrey S, Barker M, Bond L, Bonell C, Hardeman W, et al. Process evaluation of complex interventions: Medical Research Council guidance. Bmj [Internet]. 2015;350(mar19 6):h1258–h1258. Available from: http://www.bmj.com/cgi/doi/10.1136/bmj.h1258
6. De Geest S, Sabaté E. Adherence to long-term therapies: evidence for action. Eur J Cardiovasc Nurs [Internet]. 2003 Dec;2(4):323. Available from: http://www.ncbi.nlm.nih.gov/pubmed/14667488
7. McDonald HP, Garg AX, Haynes RB. Interventions to enhance patient adherence to medication prescriptions: scientific review. JAMA [Internet]. 2002 Dec 11;288(22):2868–79. Available from: http://jama.jamanetwork.com/article.aspx?doi=10.1001/jama.288.22.2868
8. Andrikopoulou E, Scott PJ, Herrera H. Mixed methods protocol for a realist evaluation of electronic personal health records design features and use to support medication adherence (ePHRma). BMJ Heal Care Informatics. 2020 Jan 5;27(1).
9. Creswell JW, Plano Clark VL. Designing and conducting mixed methods research. Los Angeles : SAGE; 2017.
10. Ritchie J, Spencer L. Qualitative data analysis for applied policy research. In: Huberman AM, Miles MB, editors. The qualitative researcher's companion [Internet]. Thousand Oaks, California: SAGE Publications, Inc.; 2002. p. 305–29. Available from: http://methods.sagepub.com/book/the-qualitative-researchers-companion
11. Pawson R. Evidence-based policy: a realist perspective. London : SAGE; 2006.
12. Michie S, van Stralen MM, West R. The behaviour change wheel: A new method for characterising and designing behaviour change interventions. Implement Sci [Internet]. 2011 Dec 23;6(1):42. Available from: http://implementationscience.biomedcentral.com/articles/10.1186/1748-5908-6-42
13. Easthall C, Barnett N. Using Theory to Explore the Determinants of Medication Adherence; Moving Away from a One-Size-Fits-All Approach. Pharmacy [Internet]. 2017 [cited 2019 Aug 28];5(4):50. Available from: www.mdpi.com/journal/pharmacy

Challenges of Trustable AI and Added-Value on Health
B. Séroussi et al. (Eds.)

doi:10.3233/SHTI220573

Patient Values Associated with an Exergame Supporting COPD Treatment

Kira OBERSCHMIDT[a,b,1], Marijke BROEKHUIS[a,b] , and Christiane GRÜNLOH[a,b]
[a] *eHealth Department, Roessingh Research and Development, The Netherlands*
[b] *Biomedical Signals and Systems Group, University of Twente, The Netherlands*

Abstract. Exercise games (exergames) can help COPD patients stay active and prevent exacerbation. When evaluating such exergames, patient values are an important variable to take into account. In this study, seven COPD patients used an exergame technology at the physiotherapist for six months. Their values regarding treatment and the exergames were identified in interviews. Values were very stable throughout the study, and closely interconnected. Personal Guidance and Independence were important values. Additionally, patients sometimes held conflicting values that they prioritized differently at different times or based on specific events. As the study identified important values that appeared stable over a period of time, albeit with different prioritization, they are worth considering when designing new technology. However, values cannot be looked at in isolation because of the strong connection between values.

Keywords. Patient values, COPD, eHealth, Value sensitive design, Exergames, Rehabilitation, Physical therapy, Value tension

1. Introduction

Chronic Obstructive Pulmonary Disease (COPD) is a common, preventable and treatable disease that is characterized by persistent respiratory symptoms and airflow limitation, a major cause for chronic morbidity and mortality and is among the top three causes of death worldwide (1). Being a chronic condition, many people suffer from COPD for years with respiratory symptoms such as dyspnea, cough and/or sputum production and periods of acute worsening of these symptoms called exacerbations (1).

It is important for COPD patients to stay active and keep exercising. Research suggests that exergames (exercise games) can be a good way of engaging COPD patients to exercise and stay active (2). When developing these games, much attention is paid to ensure safety, adherence and enjoyment (e.g. (3,4)), however, it is also important to investigate whether the technology supports the values of patients, as this might be a deciding factor whether the technology will be adopted and used. Value sensitive design (VSD) is an approach that addresses human values in the design process (5). While the word 'value' can refer to something of economic worth, VSD uses a broader meaning in which human values are defined as "what is important to people in their lives, with a focus on ethics and morality" (6) (e.g., trust, privacy, autonomy, identity). As values are contextual, may change over time or come into conflict (within an individual, among

[1] Kira Oberschmidt, Corresponding author, eHealth Cluster, Roessingh Research and Development, Postbus 310, 7500 AH Enschede, The Netherlands; E-mail: k.oberschmidt@rrd.nl.

individuals or groups, creating so-called value tensions) (6), it is important to conduct empirical investigations to examine which values are affected in the distinct setting, i.e. how they are lived and experienced in a particular context at a particular point in time. It is therefore also important to evaluate in which way the values are supported or hindered in a particular setting. This study therefore aims to answer three research questions: (1) Which values are important for COPD patients in physical therapy? (2) How are their values supported when using the exergame? (3) How and to what extend do these values change over time?

2. Method

This study was part of an AAL project (Active Assisted Living) which evaluated an exergame technology installed at a physiotherapist's office in the Netherlands, where patients perform physiotherapy exercises in a large room, sometimes in groups. Patients used the exergame as an integrated part of their treatment, not as an addition on top of the regular care. The technology consisted of a TV screen on which videos and games were shown, and a motion-sensing camera that captured the participants' movement. When detecting that a patient was carrying out an exercise incorrectly, an audio-visual notification was provided. However, sometimes these notifications erroneously occurred even though the exercises were correctly performed.

Patients were recruited by their physiotherapist and interviewed four times within a period of six months (T1 start of the project, T2 6 weeks, T3 12 weeks, T4 24 weeks after starting). In each interview, their perceived values concerning (a) their physiotherapy in general, and (b) the use of the exergame were discussed. Because the patients had not experienced the exergame at the time of the first interview, it was possible to discuss the topics of values and technology more broadly. In this first interview (T1), open questions were asked about what is important to the patients in their treatment, and with regard to technology. Based on the T1 interviews and previous research within the researchers' institution, a list of nine values that seemed important for COPD patients in rehabilitation care was created. In all following interviews (T2-T4) participants were presented with this list and asked to select a maximum of three values that were the most important to them. For each selected value, participants were asked to describe what that value means for them, how that value is reflected in their treatment (generally and specifically with the exergame) and how this could be improved.

3. Results

Seven COPD patients took part in this study, of which two completed the whole six-month study. Most dropouts occurred after the third interview (between week 12 and 24). These dropouts happened mainly due to exacerbation or co-morbidities, except for one patient who dropped out because they disliked the exergames. Participants' age ranged between 55 and 74, and five of them were male. Most had been diagnosed with COPD between four to eight years ago, and most were also following physiotherapy (group) training for as long. The GOLD status (i.e. severity of airflow limitation) of all but one patient was 2 (moderate), for the last patient this was between 3 (severe) and 4 (very severe). Based on the initial interviews and previous research conducted in the rehabilitation center, nine values were identified in relation to rehabilitation care and the

exergame that were used as prompts in the follow-up interviews: Personal Guidance, Challenge, Trust, Independence, Contact with Others, Seeing Results, Regularity, Quality of Care and Privacy.

Important values, interconnection and tensions: Patients valued INDEPENDENCE and PERSONAL GUIDANCE highly, which was shown by the fact that these values were mentioned the most and most consistently over time. PERSONAL GUIDANCE was closely related to the values QUALITY OF CARE and TRUST. Some patients mentioned that QUALITY OF CARE improved when they received PERSONAL GUIDANCE. Although TRUST was not explicitly selected, participants talked about the trust towards their therapist when being personally guided.

The two most important values, INDEPENDENCE and PERSONAL GUIDANCE, seem to be contradictory, and one patient indeed described them as "*yin and yang*", meaning that they sometimes needed a lot of support, but at other times preferred to be left alone during their treatment. One patient selected INDEPENDENCE, because "*I have difficulty asking others for help. And if I am independent [working with the exergame] I don't need to do that*". However, in the next interview, the same participant mentioned needing PERSONAL GUIDANCE while working with the exergame[2]. Two other values that were interconnected and seem contradictory are REGULARITY and CHALLENGE. Patients liked having a fixed schedule and exercise pattern but also enjoyed being tested and pushed out of their current flow from time to time.

Values that increase motivation for treatment: For some patients, CONTACT WITH OTHERS was also important. Many of the participants were part of COPD treatment groups and highly valued the contact with peers who "*know each other well*" and "*all have the same problem*". Being able to exchange experiences and feeling understood when something was not going well were the main benefits of being part of a group. However, the exergame was used by one participant at a time, and thus did not allow for interaction with peers during the session. This was especially negatively remarked upon by one participant, for whom the group was what motivated and pushed them to do their training. A value that connected to this motivating experience when being in a group was that of CHALLENGE. The participants enjoyed the competition element in the exergame and perceived it as a way to either compare themselves with other users or others in their therapy group or to challenge their own score from the previous week.

These challenges also pushed them to do better and see the effect that this had on their condition, which was reflected in the value SEEING RESULTS. Patients were motivated by seeing the effect of their treatment, especially because for some of them it was crucial to keep being active. Additionally, when seeing these positive results, they knew that they were not wasting their time. Some of the patients also mentioned REGULARITY to be important during their training. For some the regular training schedule was needed as a motivational push, because "*sometimes I don't want to, but the appointment is set and so I go*", and that in the end "*it's always nice*". Concerning the exergame in this study, this patient added that the fixed time for each game made them stick to it instead of taking a break.

Values as a barrier: PRIVACY was mentioned specifically regarding the exergame technology. Prior to the third interview with one patient, there was an issue, where a photo of another participant was accidentally uploaded to their profile by the therapist. This made the participant think about PRIVACY, because they were now concerned that

[2] This was related to the notification from the exergame about incorrectly performing an exercise, when according to the therapist, they were doing the exercise correctly.

"*somebody else can see my session*". While this was not explicitly mentioned by the participant, this situation also relates to TRUST in the system as the participant now was doubting. Additionally, this patient did not like being watched while doing the exergame and the loud notifications drew the attention of others in the room. These notifications made the patient self-conscious, even when being aware that they were often unjustly displayed.

Value change over time: Overall, the values that patients selected were relatively stable over time. The two participants who completed the study of 6 months, selected the exact same values each time, even though they were not aware of this and even surprised themselves. For another participant that dropped out after T3 the change in selected values was based mainly on the privacy issue that was described above.

4. Discussion

This study looked at important values of patients with COPD related to their treatment and the exergame that was tested. Additionally, the value change over time was investigated. Nine different values were identified and discussed by patients: Personal Guidance, Challenge, Trust, Independence, Contact with Others, Seeing Results, Regularity, Quality of Care and Privacy. Within their treatment in general, the symbiosis between PERSONAL GUIDANCE and INDEPENDENCE was deemed very important, as was CONTACT WITH OTHERS, especially others with COPD who experience the same problems. This is supported by (4) who notes that patients with COPD are often isolated due to their condition, and benefit from social contact. The exergame supported some of the patients' values (e.g. INDEPENDENCE, CHALLENGE, REGULARITY), but it did not support or even hindered other values (e.g. PERSONAL GUIDANCE, CONTACT WITH OTHERS, PRIVACY).

From the values that patients mentioned, selected and how they described them, it became apparent that a personalized approach to care is necessary. Not only can two patients have very different needs when it comes to PERSONAL GUIDANCE, but there can also be a vast difference for the same patient. Given the different situations at different times they might need a lot of support or prefer to be independent. This is supported by literature that says that suitable support from healthcare professionals increases the motivation of COPD patients to follow an exercise program (3,7). A similar personalized approach is needed when it comes to technology. Exergames support certain values but not others, so they might be more suited to a patient who for example values INDEPENDENCE highly.

Similar to (6), we also observed that values do not exist in isolation, but are interconnected. Some values are closely related and support or even depend on each other (e.g. TRUST and PRIVACY), others seem contradictory and create value tensions (e.g. PERSONAL GUIDANCE and INDEPENDENCE; REGULARITY and CHALLENGE). Most patients seemed unaware of the tension that appears between the values they chose. In addition to being very interconnected, values were also stable over time, even though the reasoning and prioritization might change, allowing patients to hold seemingly contradictory values. The interconnection between values shows that it is necessary to take a holistic approach to VSD instead of only designing for one value, e.g. INDEPENDENCE. It also shows that not all value tensions must be resolved, as people are able to balance and prioritize their values in a particular situation, especially when this flexibility is supported by the context, which is also described by (6).

A limitation of this study is the timeline of six months. To investigate the long term consistency of human values, a longer study would be needed. We did however see that, even though patients were not aware of it or consciously doing so, they selected the same values throughout the course of this study. Nonetheless, following patients for an even longer period of time could be valuable to learn more about the consistency and changes of values over time. In addition, many patients had to drop out due to exacerbations, which is something that needs to be accounted for with COPD patients. Participants who experienced exacerbations were unable to continue using the system, and in some cases also felt unable to partake in the interviews. As technology use was a main reference point for the patient values and out of respect for their wishes, the researchers did not try to convince dropouts to continue participation.

5. Conclusion

COPD patients hold a variety of different, but interconnected values. The most important values are PERSONAL GUIDANCE from their healthcare professional and keeping their INDEPENDENCE. While exergames can support some patients and patient values, others are not supported or even inhibited by technology. Even if values stay consistent over time for an individual, the value set may include values that seem to be in tension and are then prioritized given the current situation or specific important events. Generally, the values a person holds are often interconnected and relate to each other, so values should not be looked at in isolation. Therefore, when designing technology, the interconnections, dependencies and tensions between human values of stakeholders should be taken into account. This study is conducted within Active Assisted Living (AAL) project SALSA (project number: 2018-5-46-CP).

References

1. Global Initiative for Chronic Obstructive Lung Disease. Global strategy for the diagnosis, management, and prevention of chronic obstructive pulmonary disease (2021 Report) [Internet]. 2020. Available from: https://goldcopd.org/wp-content/uploads/2020/11/GOLD-REPORT-2021-v1.1-25Nov20_WMV.pdf
2. Parent A-A, Comtois A-S. Pilot project: Feasibility of high-intensity active video game with COPD patients. Tools for home rehabilitation. In: 2017 International Conference on Virtual Rehabilitation (ICVR). IEEE; 2017. p. 1–2.
3. Høeg ER, Bruun-Pedersen JR, Serafin S. Virtual Reality-Based High-Intensity Interval Training For Pulmonary Rehabilitation: A Feasibility and Acceptability Study. In: 2021 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW). IEEE; 2021. p. 242–9.
4. Wardini R, Dajczman E, Yang N, Baltzan M, Préfontaine D, Stathatos M, et al. Using a virtual game system to innovate pulmonary rehabilitation: safety, adherence and enjoyment in severe chronic obstructive pulmonary disease. Can Respir J. 2013;20(5):357–61.
5. Friedman B, Kahn Jr PH, Borning A. Value sensitive design and information systems. In: Zhang P, Galletta D, editors. Human-computer interaction management information systems: Foundations. ME Sharpe; 2006. p. 348–72.
6. Friedman B, Hendry DG. Value sensitive design: Shaping technology with moral imagination. Mit Press; 2019.
7. Sohanpal R, Steed L, Mars T, Taylor SJC. Understanding patient participation behaviour in studies of COPD support programmes such as pulmonary rehabilitation and self-management: A qualitative synthesis with application of theory. npj Prim Care Respir Med. 2015;25(June):1–15.

Challenges of Trustable AI and Added-Value on Health
B. Séroussi et al. (Eds.)

doi:10.3233/SHTI220574

Experiences Using Patient and Public Involvement in Digital Health Research for Multiple Sclerosis

Tiia YRTTIAHO[a,1] , Minna ISOMURSU[a] and Guido GIUNTI[ab]
[a] *University of Oulu, Finland*
[b] *Delft University of Technology, Delft, The Netherlands*

Abstract. Patient and public involvement (PPI) is increasingly used for improving quality of the research. There are many barriers in translating PPI into practice, including lacking examples of good practices. Frameworks that have been developed in one setting do not readily transfer to other settings. In this paper, we examine the implementation of PPI in the context of a digital health research project that explores the design, development and use of mHealth for persons with Multiple Sclerosis taking an iterative user-centered design approach. **Methods**: Instrumental case study to describe the PPI process on a digital health research project. **Results**: Overall experience was positive. We found 3 roles for PPI involvement: strategic members; design and development partners; and expert members. Challenges lay on unclear PPI terminology; managing roles and expectations; and ensuring accessibility.

Keywords. patient and public involvement, Multiple sclerosis, digital health, mhealth

1. Introduction

The promise that delivering healthcare through mobile devices (mHealth) holds relies heavily on its adoption. One of the greater challenges that digital tools face is properly meeting the users' needs, failure to do so often results in misused or underutilized solutions [1,2]. There is evidence that involvement of users in the design and development of digital solutions increases the chances that the end result is valuable and meets the needs of the users [1,2]. There have been steps towards an active involvement of patients and other stakeholders in the design process of digital health solutions, but this is still not the norm [3]. The National Institute for Health Research has defined public involvement in research as "research being carried out 'with' or 'by' members of the public rather than 'to', 'about' or 'for' them", in which public includes patients, other people and organizations that use health and social care services [4]. Patient and public involvement (PPI) is an emerging approach to increase inclusiveness in health research. The use of PPI can improve research actions since researchers "don't know what they don't know" at the beginning of their projects. This is also the reason why no one knows

[1] Corresponding author: Tiia Yrttiaho. University of Oulu. P.O. BOX 8000, FI-90014, Oulu, Finland. E-mail: tiia.yrttiaho@oulu.fi

beforehand how much more information is gained through PPI and therefore impact can seem unpredictable [5].

Patients and public have been involved many ways in different stages of research[6]. Crocker et al's [7] meta-analysis found that PPI interventions increased the odds of participant enrolment in clinical trials, especially if people involved had lived experience of the condition that was studied. PPI is a known concept especially in UK where it's also requirement for public research funding; this practice is becoming common in other parts of Europe like the Netherlands and Scandinavian countries [8].

There are many barriers in translating PPI policy into practice, including lacking examples of good practice [8]. Frameworks that have been developed in one setting do not readily transfer to other settings: there is a risk of tokenism and neglecting democratic values [6]. Successful involvement is built on equal partnership, where everyone is acknowledged, rewarded and valued [9].

Multiple Sclerosis (MS) is one of the world's most common neurologic disorders of the young adults leading to severe disability [10], and it requires significant active support. There are cases of pwMS involvement for digital health before where it has improved the quality of the research and lead to impactful changes [11,12]. In this paper, we examine the implementation of PPI in a context of a digital health research project called More Stamina. The project explores the design, development and use of mHealth for persons with Multiple Sclerosis (pwMS) taking a iterative user-centered design approach [13].

2. Methods

In this paper, we use the More Stamina project as an instrumental case study, to describe our PPI process and, based on our experiences, provide insights to better guide the use of PPI in MS mHealth projects.

More Stamina is a Research to Business project of the University of Oulu, funded by Business Finland. The team behind it is multidisciplinary, composed of physicians, psychologists, software engineers, interaction designers, information systems specialists, business developers and others.

3. Results

3.1. Patient Roles

Throughout the project, we examined how we could involve patients in all activities and phases. Through this process, the following 3 roles emerged:

- **Strategic members**. A pwMS sits on the project steering group, participating in strategic planning and decision making.
- Design and development partners. Two pwMS worked in design and development activities as members of the project team.
- Expert members. Specific roles which required professional skills involved pwMS. E.g., a User Research Professional who is a pwMS participated in interview data analysis.

3.2. PPI Process

We started PPI in this research without prior experience of PPI. Our first actual PPI activity was a test to explore possibilities of patient involvement through a pilot interview using our **strategic member**. In this session they let us know that the length was too much for them and it raised our awareness, prompting further work on this PPI line. The process began with a planning workshop among the research team, where we familiarized ourselves with the principles of PPI and discussed how it could be integrated as an essential element in our project. We nominated a person to be responsible of PPI activities and to act as contact point. Internally, we established PPI guiding principles and discussed how would we approach potential "patient representatives", what kind of reimbursement model should we use, how should PPI activities be designed, etc. A specific PPI budget and hourly fees were settled as per patient association guidelines.

We approached pwMS through a local MS association, which already had an appointed patient representative. We were able to connect with other pwMS who had prior experience working as patient representatives and they were invited to join the project. A second patient representative was found through the first's extended network. A total of 2 patient representatives were recruited as **design and development partners**.

We designed an onboarding process where the first meeting was reserved for mutual learning and getting to know each other. The research team presented the project scope, previous research and our partner institutions; pwMS gave us context on their history with MS. We used this meeting to inquire the patient representatives about their preferences and validate types of tasks. To reduce the risk of overburdening, an initial estimation of PPI activities was agreed: monthly meetings, with a maximum session length of 2 hours. There was also a clear preference of keeping online sessions short and leave more time for face-to-face encounters. PPI activities were determined to consist of hands-on workshops and occasional consultation. COVID-19 restrictions played a role so some in-person sessions had to be rescheduled as online meetings. In order to keep our patient representatives up to date when no active tasks were required, they were added to project's monthly newsletter and internal mailing list.

In-person meeting were held in the premises of University of Oulu, as it has accessible entries, toilets, parking, and good public transport. It was checked beforehand that meeting rooms were accessible, and we mapped unobstructed routes so we could guide them. This map was sent to patient representatives before each meeting.

In the 5 months that patient representatives have been involved in our project, we have consulted them on several instances. For example, in a task about mapping the MS patient journey through qualitative interviews to healthcare professionals and pwMS. **Design and development partners** went over the preliminary materials to be used in research, the pwMS recruitment plan, and the creation of the More Stamina tutorials that would be beneficial to participants in our research. Later, a pwMS acted as **expert member**, going over the collected data and providing insights.

3.3. Challenges in PPI for Digital Health Research

3.3.1. Unclear Terminology

Early in our process it became clear that different terms have been used interchangeably although they are not such similar concepts: consumer involvement, patient engagement or co-production. It was often difficult for pwMS to differentiate between PPI and

traditional design activities, such as usability testing. Terminology got further complicated because of the multilingual nature of the team and how PPI terminology differs between languages.

3.3.2. Roles and Expectations

Even though in our project we tried to be clear and open with the patient representatives, it still left us moments of confusion. It has been a challenge to inform our patient representatives and keep their roles and expectations clear, while at the same time avoid confusing them with unnecessary information.

The PPI process has been new to our team, so roles also have been defined as we have worked together. Our main role in PPI has been supporting and empowering patient representatives and creating collaborative atmosphere so they are able to contribute. We have been listening to their hopes and asked for feedback.

3.3.3. Accessibility

MS-disease can cause changes in motor and cognitive functions, and it can also cause disabling fatigue. Keeping this in mind, we invested extra time in removing accessibility barriers (e.g., ensuring wheelchair access and considering session length). Accessibility is more than just accommodating for physical limitations. In our case, not only information had to be accessible, but also technology. Any use of digital tools during remote sessions took extra planning to minimize the burden a tutorial could become.

4. Discussion

Successful PPI lies deeply on a shared understanding and power balance [9,14]. It is important to generate an atmosphere of equality among the team, which may not be easy to accomplish, as healthcare contexts carry inherent power-asymmetries. This is something that in digital health projects can be compounded with the added potential disadvantage of digital literacy. PPI literature indicates how remuneration is useful not only as means of compensation, but also to even out possible power difference between patient representatives and the rest of the research team [6].

Terminology issues might be slowing the spread of PPI good practices. NIHR remarks that involvement, engagement and participation are used interchangeably and gives them definition that differs from involvement [4]. Biddle et al [8] also noted that there was inconsistency in terms and purpose of PPI.

Unclear information and lack of preparation has been reported to impair patient representatives ability to contribute [15]. Our information had to be understandable also for people without healthcare or research background, so it was important avoid professional jargon and too technical language [14].

The literature highlights the importance of clarifying the expected types of tasks and time commitment [14]. Clarity of roles is considered important in PPI [14], yet it seems to be still a common issue [14,15]. Active involvement in research activities can be a burden for patient representatives [15], and this was even more sensitive for MS [12]. It was interesting to observe that pwMS were very skilled in estimating how much time they would be able to contribute. When this was clearly agreed, they felt it was a fair situation.

A common complaint of PPI in research is that it takes more time [15]. Setting up a project's PPI process certainly requires time, knowledge, and resources. However, we think that using this time early in the process was beneficial by giving us a better understanding of needs and the potential value of the solution.

5. Conclusions

Our positive experience is aligned with previous literature. Valuable insight was gained from PPI that guided research actions. Even in situations where our patient representatives had nothing to add, it helped in validating our approach. The process clarified the use of PPI, which will make further PPI actions more efficient in the future.

References

[1] Schnall R, Rojas M, Bakken S, et al. A user-centered model for designing consumer mobile health (mHealth) applications (apps). *J Biomed Inform*. 2016 60:243-251.
[2] Brown W, Yen PY, Rojas M, Schnall R. Assessment of the Health IT Usability Evaluation Model (Health-ITUEM) for evaluating mobile health (mHealth) technology. *J Biomed Inform*. 2013 46(6):1080-1087.
[3] Mirkovic J, Jessen S, Kristjansdottir OB, Krogseth T, Koricho AT, Ruland CM. Developing Technology to Mobilize Personal Strengths in People with Chronic Illness: Positive Codesign Approach. *JMIR Form Res*. 2018 2(1):e10774.
[4] National Institute for Health Research. Briefing notes for researchers - public involvement in NHS, health and social care research. Published 2021. https://www.nihr.ac.uk/documents/briefing-notes-for-researchers-public-involvement-in-nhs-health-and-social-care-research/27371
[5] Staley K. 'Is it worth doing?' Measuring the impact of patient and public involvement in research. *Res Involv Engagem*. 2015 1(1):6.
[6] Greenhalgh T, Hinton L, Finlay T, et al. Frameworks for supporting patient and public involvement in research: Systematic review and co-design pilot. *Heal Expect*. 2019 22(4):785-801.
[7] Crocker JC, Ricci-Cabello I, Parker A, et al. Impact of patient and public involvement on enrolment and retention in clinical trials: systematic review and meta-analysis. *BMJ*. 2018 363:k4738.
[8] Biddle MSY, Gibson A, Evans D. Attitudes and approaches to patient and public involvement across Europe: A systematic review. *Health Soc Care Community*. 2021 29(1):18-27.
[9] Baines RL, Regan de Bere S. Optimizing patient and public involvement (PPI): Identifying its "essential" and "desirable" principles using a systematic review and modified Delphi methodology. *Heal Expect*. 2018 21(1):327-335.
[10] Goldenberg MM. Multiple sclerosis review. *P T*. 2012 37(3):175-184. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3351877&tool=pmcentrez&rendertype=abstract
[11] Allen-Philbey K, Middleton R, Tuite-Dalton K, et al. Can We Improve the Monitoring of People With Multiple Sclerosis Using Simple Tools, Data Sharing, and Patient Engagement? *Front Neurol*. 2020 June;11:1-7.
[12] Synnot AJ, Cherry CL, Summers MP, et al. Consumer engagement critical to success in an Australian research project: reflections from those involved. *Aust J Prim Health*. 2018 24(3):197.
[13] Giunti G, Mylonopoulou V, Rivera Romero O. More Stamina, a Gamified mHealth Solution for Persons with Multiple Sclerosis: Research Through Design. *JMIR mHealth uHealth*. 2018 6(3):e51.
[14] Ocloo J, Garfield S, Franklin BD, Dawson S. Exploring the theory, barriers and enablers for patient and public involvement across health, social care and patient safety: a systematic review of reviews. *Heal Res Policy Syst*. 2021 19(1):8.
[15] Brett J, Staniszewska S, Mockford C, et al. A Systematic Review of the Impact of Patient and Public Involvement on Service Users, Researchers and Communities. *Patient - Patient-Centered Outcomes Res*. 2014 7(4):387-395.

Challenges of Trustable AI and Added-Value on Health
B. Séroussi et al. (Eds.)

doi:10.3233/SHTI220575

Mindful Workarounds in Bar Code Medication Administration

Valentina LICHTNER[a,1] and Dawn DOWDING[b]
[a] *Leeds University Business School, University of Leeds, UK*
[b] *School of Health Sciences, University of Manchester, UK*

Abstract. Bar-Coded Medication Administration systems (BCMA) are often used with workarounds. These workarounds are usually judged against standard operating procedures or the use of the technology as 'designers' intended'. However, some workarounds may be reasonable and justified to prevent safety errors. In this conceptual paper, we clarify BCMA safety mechanisms and provide a framework to identify workarounds with BCMA that nullify the error prevention mechanisms inherent in the technology design and process. We also highlight the importance of understanding the purpose behind a nurse's workaround in BCMA, focusing on the notion of mindful (thoughtful) workarounds that have the potential to improve patient safety.

Keywords. Medication safety, workarounds, barcoding, closed-loop systems

1. Introduction

Bar-Coded Medication Administration systems (BCMA) are implemented in hospitals to reduce preventable medication errors and improve medication safety [1, 2]. Research suggests that for BCMA to contribute to safety, BCMA must be used as designers' intended, without workarounds, as these may lead to medication administration errors [1, 3, 4]. Nurses are the main users of BCMA and also considered the last line of defense in preventing medication errors reaching the patient.

However, research in high-risk organizations has shown how safety is sometimes achieved by operators not using technology as designers intended [8]. In circumstances of danger unplanned for by designers, staff resourcefulness and deviations from protocols are mechanisms for resilience and contain potential harm. This contribution to safety should be considered when approaching the implementation of BCMA technology and assessing its impact. Some violations to BCMA procedures may be due to nurses' 'good judgement' [5] and enable safe care. It is therefore necessary to distinguish between workarounds of different quality – those that introduce risks to patient safety, those that do not introduce risks, and those that prevent or contain risks (also potentially created by the technology). An understanding of the mechanisms conducive to safety (or errors) may inform this distinction.

In this conceptual paper, we wish to distinguish between workarounds with BCMA that nullify the technology's safety mechanisms, and those that do not. We believe this would lead to a better understanding of the contribution of BCMA to medication safety.

[1] Leeds University Business School, University of Leeds, Leeds LS2 9JT, UK - v.lichtner@leeds.ac.uk

2. Methods

We sought to identify the mechanisms that may be conducive of safety in nurses' administration of medications to hospital patients with BCMA [6, 7].

We analyzed conceptually the typical BCMA functionalities to verify that the right medication (including right dose, route, and time) is being administered to the right patient, by scanning barcodes of both the patient and the medications against an electronic record. With such a system, the patient is associated to a machine readable barcode, typically by wearing a wristband; the medication's packaging (possibly in single dose) also bears a barcode that identifies the item. We analyzed and modelled the intended high-level workflow of such a BCMA (Figure 1) to identify the mechanisms of error prevention and underlying assumptions. We then hypothesized how workarounds in BCMA use could nullify (or not) each of the mechanisms we identified and how errors may then occur.

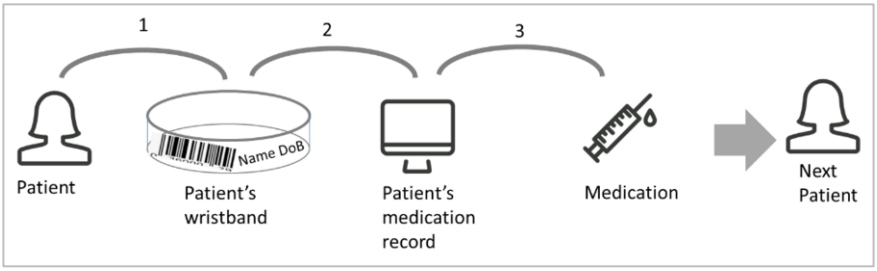


Figure 1. Bar-Coded Medication Administration (BCMA) high-level workflow for error prevention

3. Results

3.1. Mechanisms of error prevention

BCMA prevents medication errors by a three way match verification process involving four steps, performed at each administration, possibly in any order:

I. Verification that the correct barcode is associated with the intended patient, e.g. by asking the patient to confirm their name and date of birth, and matching this information with patient details indicated on the wristband *(right patient)*.

II. Verification that the correct medication record has been selected, by scanning the patient barcode with the electronic medication system *(right record)*. The system automatically opens the record, or alerts the nurse if a record cannot be found.

III. Verification that the correct medication has been prepared for administration, as prescribed in the patient record, by scanning the medication barcode with the electronic medication system. The system alerts the nurse if the medication does not match what has been prescribed *(right medication – item, dose, time, route)*.

IV. The three verifications steps are performed for each patient and their medications separately, not in batches over multiple patients. Nurses will prepare, verify and give medications to each patient one at a time. The risk

otherwise is that the medications may be mixed-up between patients *(right patient, right medication)*.

In addition, only if all three verification steps have been completed before (and not after) administration of the medication for each single patient (not in batches), the system may prevent administration errors. It is also necessary that the nurse does not disregard any alerts generated in these verification steps [8]. However, for safety to be achieved, this set of mechanisms also assumes that the right medications have been prescribed to the patient, in the correct patient record; otherwise the wrong (unintended) medication may be administered to the (unintended) patient.

3.2. Nullifying and not-nullifying workarounds

We propose that workarounds that nullify BCMA's error prevention mechanisms are those where only some, but not all, of the steps are completed. For example, when the nurse assumes that the barcode on the patient's wristband is correct and does not complete step I in the process; when the nurse does not use the patient wristband to open the record (step II), and instead selects the record on screen; when the nurse scans the medications after (instead of before) giving it to the patient; when the nurse scans multiple patients, and/or medications intended for different patients at the same time (processing the task in batches – IV). It is also possible that the steps are completed but without paying attention to alerts [8] or to 'what happened on the computer' [9] at the time of scanning barcodes.

Instead, not-nullifying workarounds are those where nurses adapt the use of BCMA to contextual aspects of the medication administration process, but still enable the nurse to complete all verification steps. For example, the patient wristband may not be physically on the patient, but the nurse still checks with the patient that the barcode on the wristband is correct (step I). These workarounds are not nullifying the error prevention mechanisms of the BCMA as intended by its design, and suggest that the nurse is 'mindful' (thoughtful) in their approach to the use of BCMA.

3.3. When the underlying assumptions do not hold true – the wrong order

A major risk introduced by the use of BCMA is the medication being given to an unintended patient when the order has been entered in the wrong patient record.

If doctors have prescribed a medication in the wrong record, BCMA may not prevent administration, or alert the nurse of the error. This was noted in a study [10], where a nurse explained: *'I know my patient and then an order will pop up, "Give Lopressor or give whatever to this patient" and I know that can't be for my patient. Sometimes physicians inadvertently make a wrong order on a patient'* (p85).

Instead, BCMA introduces the risk that nurses will rely on the technology and give to the patient what has been ordered without questioning it, as the technology 'says so' (non-mindful use of the BCMA). The main barrier to such errors is the nurse's quality of attention, as well as their knowledge of the patients, highlighting again the importance of 'mindful' approaches to the use of BCMA. In the context of reliability and safety, mindfulness indicates a critical stance [11, 12], such as nurses' asking questions of the technology and of the prescription.

Different combinations of nullifying and not-nullifying workarounds, and their relationship to 'mindful' use of the BCMA are summarized in Table 1.

Table 1. Examples of safety mechanisms of barcoded medication administration (BCMA), the nullifying/not nullifying effects of workarounds in combination across the different mechanisms and nurses' mindful (thoughtful) or non-mindful (over-reliance on technology) use of BCMA

Effect of combination of workarounds	Safety mechanisms				Description	Mindful use of BCMA
	I	II	III	IV		
Not-nullifying	Y	Y	Y	Y	Despite BCMA not used to the rule, nurses' actions still enabled to confirm right patient, right record and right medication	Mindful
Nullifying	Y	N	?	?	Despite BCMA use, and confirmation of the right patient, there has been no check on whether the record is right for this patient, thus a risk the medications are not right	Non-Mindful
Nullifying	N	?	?	?	Despite BCMA use, and a match on the record, there has been no check on whether the patient barcode is right, thus a risk that record and medications are not right	Non-Mindful
Nullifying	Y	Y	N	?	Despite BCMA use, and confirmation of the right patient, and check that the record is right for this patient, there has been no check on whether the medications are right	Non-Mindful
Nullifying	Y	Y	Y	N	Despite BCMA use, by processing medications across patients in batches there is a risk of medication mix-ups	Non-Mindful
No workarounds	Y	Y	Y	Y	Compliance with BCMA intended use may not detect orders written in the wrong record (unless mindful use)	Mindful/Non-mindful

Note: I = verifying patient with patient barcode; II = matching medication record with patient barcode; III = matching medication barcode with medication record; IV = verification one patient at a time, and with attention to alerts; Y/N = yes/no to performing each verification step; ? = either Y or N. The table does not include all possible alternatives; it does not intend to suggest a direct correspondence between (not-)nullifying workarounds and mindful attention; e.g. nullifying workarounds by nurses mindfully asking questions of the prescription are possible.

4. Discussion and conclusion

Workarounds are a difference between work as done and work as prescribed or imagined, explainable by a variety of sociotechnical factors (e.g. [13, 14]). In use of clinical systems, 'dangerous workarounds exist that should be reduced to improve safety' [7]. In using BCMA to reduce medication errors, the question is which workarounds are dangerous and should be reduced, and which ones contribute to safety. To answer this question, we propose that workarounds may be 'nullifying' and 'not nullifying' BCMA safety mechanisms, and nurses' attention with BCMA may be of mindful or mindless quality. Given that BCMA systems have been implemented across health care systems worldwide; having a more detailed and nuanced analysis of the types of workarounds used by nurses, including where they may actually be justified to enhance patient safety is crucial for our future understanding of technology development and usability.

We believe that understanding BCMA technology's safety mechanisms, may inform interventions to improve the safety of BCMA use. The set of mechanisms we have

identified can provide a framework for evaluating the adoption/use of BCMA. The framework may also be part of nurses' training in the use of BCMA. It is essential that nurses understand how and why this technology may improve safety, including its limitations, as understanding *why* may encourage mindful attention in using BCMA, and prevent them engaging in nullifying workarounds. Future research may empirically investigate what are the most common types of workarounds for each mechanism, which ones are nullifying/not-nullifying, what are the contexts that may influence these workarounds, and propose interventions focused on nullifying workarounds.

To conclude, we stress that in using BCMA, what is important is not necessarily compliance with designers' intentions (i.e. absence of workarounds), but the 'quality of workarounds' and the quality of attention nurses pay to the activity and the technology. There is a real risk of nurses' 'reduced critical thinking and [reduced] situational awareness' with BCMA [15]. The potential of nurses' over-reliance on technology is one of the most difficult challenges to address in pursuing medication safety.

References

[1] Strudwick G, Reisdorfer E, Warnock C, Kalia K, Sulkers H, Clark C, and Booth R, Factors Associated With Barcode Medication Administration Technology That Contribute to Patient Safety: An Integrative Review, J Nurs Care Qual 33 (2018), 79-85.
[2] Hutton K, Ding Q, and Wellman G, The Effects of Bar-coding Technology on Medication Errors: A Systematic Literature Review, J Patient Saf 17 (2021), e192-e206.
[3] van der Veen W, van den Bemt PM, Wouters H, Bates DW, Twisk JW, de Gier JJ, Taxis K, Group BS, Duyvendak M, and Luttikhuis KO, Association between workarounds and medication administration errors in bar-code-assisted medication administration in hospitals, Journal of the American Medical Informatics Association 25 (2017), 385-392.
[4] Koppel R, Wetterneck T, Telles JL, and Karsh BT, Workarounds to barcode medication administration systems: their occurrences, causes, and threats to patient safety, Journal of the American Medical Informatics Association 15 (2008), 408-423.
[5] Koppel R, Bar-coded medication administration brings trade-offs even with humane automation and lean approaches, Joint Commission Journal on Quality & Patient Safety 40 (2014), 339-340.
[6] Holden RJ, Rivera-Rodriguez AJ, Faye H, Scanlon MC, and Karsh B-T, Automation and adaptation: nurses' problem-solving behavior following the implementation of bar-coded medication administration technology, Cognition, Technology & Work 15 (2013), 283-296.
[7] Patterson ES, Workarounds to intended use of health information technology: a narrative review of the human factors engineering literature, Human Factors 60 (2018), 281-292.
[8] Sloss EA and Jones TL, Alert Types and Frequencies During Bar Code–Assisted Medication Administration: A Systematic Review, J Nurs Care Qual 35 (2020), 265-269.
[9] Hong JY, Ivory CH, VanHouten CB, Simpson CL, and Novak LL, Disappearing expertise in clinical automation: Barcode medication administration and nurse autonomy, Journal of the American Medical Informatics Association 28 (2021), 232-238.
[10] Eisenhauer LA, Hurley AC, and Dolan N, Nurses' reported thinking during medication administration, J Nurs Scholarsh 39 (2007), 82-87.
[11] Langer EJ, Mindfulness, Da Capo Press, 2014.
[12] Weick KE, Sutcliffe KM, and Obstfeld D, Organizing for High Reliability: Processes of Collective Mindfulness, Research in Organizational Behavior (1999), 81-123.
[13] Azad B and King N, Enacting computer workaround practices within a medication dispensing system, European Journal of Information Systems 17 (2008), 264-278.
[14] Blijleven V, Koelemeijer K, Wetzels M, and Jaspers M, Workarounds emerging from electronic health record system usage: consequences for patient safety, effectiveness of care, and efficiency of care, JMIR human factors 4 (2017), e7978.
[15] Williams R, Aldakhil R, Blandford A, and Jani Y, Interdisciplinary systematic review: does alignment between system and design shape adoption and use of barcode medication administration technology?, BMJ Open 11 (2021), e044419.

Challenges of Trustable AI and Added-Value on Health
B. Séroussi et al. (Eds.)

doi:10.3233/SHTI220576

How Does Mental Workload Influence the Adoption of Clinical Information Systems: An Exploratory Study

Lisanne KREMER[a,1], Ann-Kathrin SCHWARZ[a], Rainer RÖHRIG[b], Bernhard BREIL[a]
[a] *Faculty of Health Care, Niederrhein University of Applied Sciences, Krefeld, Germany*
[b] *Institute of Medical Informatics, Medical Faculty of RWTH Aachen University, Aachen, Germany*

Abstract. Mental workload and technology acceptance are relevant factors that relate to use behavior and performance. Studies show a potential moderating effect of mental workload on predictors of technology acceptance. Aim of this study was the investigation of predictors of technology acceptance (UTAUT) related to clinical information systems and their relation to mental workload. This quasi-experimental study with 48 participants used the following measures: NASA TLX and UTAUT questionnaire. Participants had to perform three tasks on a clinical information system as well as four task-levels of the n-back task with increasing difficulty. Analyses show a high level of technology acceptance (M=3.82, SD=.76) and confirm performance expectancy as the most relevant predictor of behavioral intention (β=.48, p<.001). A linear regression showed that a high level of mental workload has an influence on performance expectancy (F1,46=8.438, p<.05). The study shows an influence of mental workload on acceptance, the strength and role of which (e.g. moderation) needs to be further investigated, especially in the context of other determinants.

Keywords. Mental Workload, UTAUT, Technology Acceptance, Clinical Information Systems

1. Introduction

1.1. Background

In recent years, the mental workload (MWL) of employees in the healthcare sector has increased significantly [1]. MWL is a multimodal, multidimensional, and complex concept that describes the relationship between a person's available and therefore limited resources and the demands of a task [2]. One method to measure MWL reliably and validly is the NASA TLX [3]. The level of MWL has an influence on different areas of professional life and is a key factor when it comes to adoption of Clinical Information Systems. MWL affects general performance parameters such as decision-making behavior [4], performance [5], occupational parameters such as job stress or job satisfaction, or even mental health (increase in depression, burnout, etc.) [7]. MWL also

[1] Lisanne Kremer, Niederrhein University of Applied Sciences, Faculty of Health Care, Reinarzstr. 49, 47805 Krefeld, Germany; E-mail: Lisanne.kremer@hs-niederrhein.de

affects people who interact with technology in their daily work, which places high cognitive demands on users. The negative correlation between MWL and performance is one example of such an effect.

Beside high workloads that influence the use of and satisfaction with digital health technologies i.e. clinical information systems, the adoption of health information technology plays an important role in the efficient use of these systems [8].

One approach to investigate the adoption of healthcare technologies is the unified theory of acceptance and use of technology (UTAUT) [9]. The predictors of behavioral intention to use a technology (BI), used as a measure of acceptance, are performance expectancy (PE), effort expectancy (EE), social influence (SI) and facilitating conditions (FC) [9]. Brown identified different antecedents of EE and PE that relate to tasks and personal factors [10]. Identified antecedents include technostress as well as MWL [11,12]. Dang assumes that: an increased workload leads to a low perceived level of task performance and therefore to a low level of PE and EE when using social media search systems [12].

1.2. Aim/ Hypotheses

The aim of this explorative study was to investigate the relationship of MWL and predictors of technology acceptance/BI in context of clinical information systems
Hypotheses:

- PE, EE, FC, SI predict BI in context of using clinical information systems
- The level of MWL influences the PE/EE/BI

2. Methods

2.1. Participants

All participants were recruited from Social Media (LinkedIn) primarily at the Faculty of Health Care of the Niederrhein University of Applied Sciences. Inclusion criteria were: Aged between 18-70 years, experience in a healthcare profession or studying a healthcare related subject (e. g. nursing). The study was performed in accordance with the ethical standards laid down in the Declaration of Helsinki and approved by the ethics committee of the RWTH Aachen University (Vote-No: EK 138/21, chairman: Prof. Dr. Schmalzing). Prior to the start of the study, informed consent was obtained from each participant.

2.2. Material

A diagnostic screen, mouse and keyboard were placed on a table. The experiment consisted of two different task paradigms and three different questionnaires. The entire experimental procedure was controlled with a self-developed program in a lab [13].

The n-back-task: During the n-back task [14] the participants are shown letters on a screen and then have to decide according to a given rule whether the letter is a target or a distractor. Relating to the level of the n-back task, participants have to decide if the stimulus (letter) matches a stimulus (letter) n-trials before [15]. For the n-back task of level 1 (n=1-back task), for example, the target would be the identical letter to the previously displayed letter. The participants should respond to these targets as quickly

as possible by pressing j (for target) or k (for distractor) on the keyboard. The rule (f. ex. n=1 back) is kept constant during one segment of the experiment. By increasing n in the n-back task, we increased the difficulty of the task in subsequent segments.

Tasks performed on the PACS: The participants had to perform three tasks with increasing difficulty on a picture archiving and communication system (PACS) (©Visus) [16]. The difficulty levels were validated by three external experts (physicians, medical informatics specialists).

- Level 1: Accessing patient data (date of birth, study date) and making entries
- Level 2: Determining preliminary findings and interpreting results
- Level 3: Screening these findings and measuring an abnormality

Questionnaires: Participants were asked to answer the Raw NASA TLX [17], an adapted UTAUT questionnaire [18], and demographic questions referring to their age, qualification and job as well as their previous experience in clinical information systems/JiveX. The raw NASA TLX provides a valid measurement of the overall workload of a task and more efficient version as the original NASA TLX [19] and consists of six predefined dimensions [20,21]. Three of them measure demands (mental, physical, temporal demands), the other three measure the way a participant deals with a task (self-rated performance, effort, frustration level). Since our task did not involve physical demands, we excluded this dimension. We further used a modified version of the UTAUT questionnaire to assess overall acceptance [18] that includes PE, EE, SI and FC as predictors of BI. The questionnaire consists of 18 items rated on a 5-point Likert scale from “do not agree at all” to “fully agree”.

2.3. Experimental Design and Procedure

The experimental design and procedure are displayed in figure 1.

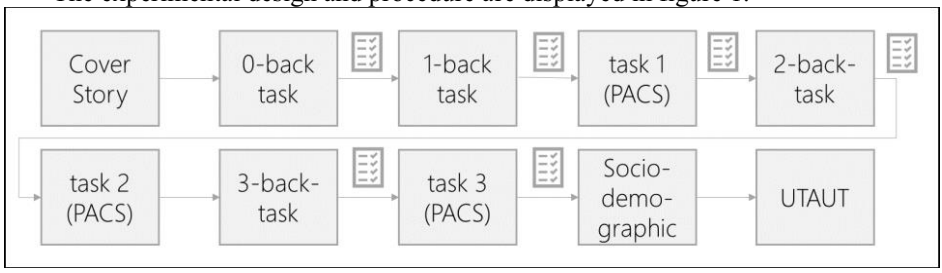


Figure 1. The experimental design should lead to an increasing workload level of the participants variated by different factors (n-back task and tasks performed on the system alternated). Each task was followed by the NASA-TLX ( ). Acceptance and sociodemographics were surveyed at the end of the implementation.

2.4. Data Analytics/Statistical Analyses

The data analysis was completed using SPSS version 27 (IBM Analytics). We processed a stepwise hierarchical regression analysis to assess the predictors of acceptance and linear regressions to assess the relationship of MWL and predictors of acceptance.

3. Results

3.1. Sample

A total of 48 subjects (66.6% female) with a mean age of 31.13 (SD 11.7) participated in the study. 11 had completed training in a health-related profession, 13 in a related field and 20 were students in health-related studies.

3.2. Reliability Analyses and Technology Acceptance of the PACS and relationship between MWL and UTAUT

The internal consistency of the subscales is listed in the following: PE (.901), EE (.791), FC (.652), SI (.921), BI (.802), NASATLX (.774 -.864). Therefore, internal consistency is acceptable to satisfying for all subscales except for the subscale of SI.

Acceptance of using clinical information systems was high (M=3.82, SD=.76, range 1-5 4, 1-5) [22]. Preliminary correlational analysis showed the highest correlations with acceptance for PE (r=.427, p<.001), SI (r=.410, p<.001), FC (r=. 411, p<.001). A significant hierarchical stepwise regression model included only 1 of 3 eligible variables ($F_{1,46}$=15.929, p<.001). There was no sign of severe multicollinearity. Only PE remained significant (β=.48, p<.001) as a predictor of BI. The explained variance was 24.1%.

For the analyses of the relationship and predictors of MWL, we proceeded with preliminary correlational analyses. There was a significant negative correlation between EE (r=-.307, p<.001), PE (r=-.319, p<.001) and high scores of the NASA TLX from tasks performed at the clinical information system.

There was a significant correlation between PE (r=.394, p<.001), SI (r=.353, p<.05) and high scores of the NASA TLX of the n-back task. A linear regression showed that a high level of MWL has an influence on PE ($F_{1,46}$=8.438, p<.05).

4. Discussion

Our study identified PE as the most relevant predictor of BI, which is in line with other findings [18]. We also identified preliminary evidence for the role of mental workload in the relationship of predictors and acceptance itself.

The individual calculations show that the studied factors have a relevant influence. Unfortunately, the sample is too small to investigate the influence of the various determinants in a multidimensional regression model.

5. Conclusions

The study shows an influence of mental workload on acceptance, the strength and role of which (e.g. moderation) needs to be further investigated, especially in the context of other determinants.

References

1. Körber M, Schmid K, Drexler H, Kiesel J. Subjective Workload, Job Satisfaction, and Work-Life-Balance of Physicians and Nurses in a Municipal Hospital in a Rural Area Compared to an Urban University Hospital. [Subjective Workload, Job Satisfaction, and Work-Life-Balance of Physicians and Nurses in a Municipal Hospital in a Rural Area Compared to an Urban University Hospital]. Gesundheitswesen 2018;80(5):444-452. PMID:28201830
2. Wickens CD. Multiple resources and mental workload. Hum Factors 2008;50(3):449-455. PMID:18689052
3. Le Hart SG S. Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research. Advances in psychology 1988(52):139-183.
4. Soria-Oliver M, López JS, Torrano F. Relations between mental workload and decision-making in an organizational setting. Psicol Reflex Crit 2017;30(1):7. PMID:32026983
5. Hancock PA, Matthews G. Workload and Performance: Associations, Insensitivities, and Dissociations. Hum Factors 2019;61(3):374-392. PMID:30521400
6. Pourteimour S, Yaghmaei S, Babamohamadi H. The relationship between mental workload and job performance among Iranian nurses providing care to COVID-19 patients: A cross-sectional study. J Nurs Manag 2021;29(6):1723-1732. PMID:33690932
7. Akca M, Küçükoğlu MT. Relationships Between Mental Workload, Burnout, and Job Performance. In: Chandan H, Realyvásquez-Vargas A, Arredondo-Soto KC, Hernández-Escobedo G, González-Reséndiz J, editors. Evaluating Mental Workload for Improved Workplace Performance: IGI Global; 2020. ISBN:9781799810520. p. 49–68.
8. Momani AM, Jamous MM, Hilles SMS. Technology Acceptance Theories. International Journal of Cyber Behavior, Psychology and Learning 2017;7(2):1-14. doi:10.4018/IJCBPL.2017040101
9. Venkatesh, Morris, Davis. User Acceptance of Information Technology: Toward a Unified View. MIS Quarterly 2003;27(3):425. doi:10.2307/30036540
10. Brown SA, Dennis AR, Venkatesh V. Predicting Collaboration Technology Use: Integrating Technology Adoption and Collaboration Research. Journal of Management Information Systems 2010;27(2):9-54. doi:10.2753/MIS0742-1222270201
11. Abd Aziz NN, Aziz MA, Abd Rahman NAS. The mediating effects of student satisfaction on technostress–performance expectancy relationship in university students. JARHE 2022. doi:10.1108/JARHE-03-2021-0117
12. Dang Y, Zhang Y, Brown SA, Chen H. Examining the impacts of mental workload and task-technology fit on user acceptance of the social media search system. Inf Syst Front 2020;22(3):697-718. doi:10.1007/s10796-018-9879-y
13. lab.js: A free, open, online experiment builder: Zenodo; 2021.
14. Kirchner WK. Age differences in short-term retention of rapidly changing informaion. Journal of Experimental Psychology 1958(55).
15. Herff C, Heger D, Fortmann O, Hennrich J, Putze F, Schultz T. Mental workload during n-back task-quantified in the prefrontal cortex using fNIRS. Front Hum Neurosci 2013;7:935. PMID:24474913
16. Visus (2017). JiveX.
17. Said S, Gozdzik M, Roche TR, Braun J, Rössler J, Kaserer A, Spahn DR, Nöthiger CB, Tscholl DW. Validation of the Raw National Aeronautics and Space Administration Task Load Index (NASA-TLX) Questionnaire to Assess Perceived Workload in Patient Monitoring Tasks: Pooled Analysis Study Using Mixed Models. J Med Internet Res 2020;22(9):e19472. PMID:32780712
18. Breil B, Salewski C, Apolinário-Hagen J. Comparing the Acceptance of Mobile Hypertension Apps for Disease Management Among Patients Versus Clinical Use Among Physicians: Cross-sectional Survey. JMIR Cardio 2022;6(1):e31617. PMID:34989683
19. Hendy KC, Hamilton KM, Landry LN. Measuring Subjective Workload: When is one scale better than many? Human Factors: The Journal of the Human Factors and Ergenomics Society 1993;35(4):579-601.
20. Hart SG. Nasa-Task Load Index (NASA-TLX); 20 Years Later. Proceedings of the Human Factors and Ergonomics Society Annual Meeting 2006;50(9):904-908. doi:10.1177/154193120605000909
21. Grier RA. How High is High? A Meta-Analysis of NASA-TLX Global Workload Scores. Proceedings of the Human Factors and Ergonomics Society Annual Meeting 2015;59(1):1727-1731. doi:10.1177/1541931215591373
22. Hennemann S, Beutel ME, Zwerenz R. Drivers and Barriers to Acceptance of Web-Based Aftercare of Patients in Inpatient Routine Care: A Cross-Sectional Survey. J Med Internet Res 2016;18(12):e337. PMID:28011445

Challenges of Trustable AI and Added-Value on Health
B. Séroussi et al. (Eds.)

doi:10.3233/SHTI220577

Health Care Professionals´ Perspectives on the Uses of Patient-Generated Health Data

Sharon GUARDADO[a,1], Minna ISOMURSU[a] and Guido GIUNTI[a,b]
[a] *Faculty of Information Technology and Electrical Engineering, University of Oulu, Oulu, Finland*
[b] *Faculty of Industrial Design Engineering, Delft University of Technology, Delft, The Netherlands*

Abstract. Integration of digital self-management solutions into health care processes requires the involvement of health care professionals in the adoption and use of the solutions as part of the care pathway. We conducted 23 interviews with diverse profiles of health care professionals participating in the treatment of chronic patients in three different countries. Our results indicate that health care professionals appeared relatively motivated at the prospect of having access to patient-generated data. Nevertheless, they appeared less confident in weighing what types of data could be collected efficiently through mobile devices and how it could be presented in ways that would provide value to the care process. Our results identify four broad categories for how patient-generated health data could be useful: monitoring, prevention, research, and transparency of condition parameters.

Keywords. Self-management mHealth, Patient-generated health data

1. Introduction

As mobile technologies improve rapidly to provide more accurate measurements of physical status and physiologic parameters, they are starting to influence different aspects of health care practices [1] around the world. The use of mobile solutions for self-management is a growing trend and has been deemed as one of the main success factors in mHealth interventions [2]. This trend, combined with the appropriate use of patient-generated health data (PGHD), could become particularly useful in clinical activities such as monitoring, treatment, and follow-up, especially in the case of patients living with chronic conditions.

PGHD is the data created, recorded, or gathered by and from patients, often through the use of mobile technologies [3]. In patient-centered management models, this type of data can be useful for disease management and prevention[1], and related activities such as motivating patients to adhere to treatment or guiding them into healthier lifestyles [4]. However, several challenges arise as massive amounts of mHealth PGHD from wearable and mobile devices come into the scene, requiring efficient analysis methods and approaches to process and evaluate the collected data, to design actual solutions that could assist HCP and patients in shared decision-making is required. Previous research

[1] Corresponding Author, Sharon Guardado, Faculty of Information Technology and Electrical Engineering, University of Oulu, Pentti Kaiteran katu 1, 90570 Oulu, Finland; E-mail: sharon.guardadomedina@oulu.fi.

has shown that for HCP, involving medical experts in the design and development of digital health technologies is a determinant aspect for their endorsement and support [5]. Similarly, for most HCP, a solid scientific foundation and theory backing digital solutions is a prerequisite for their validation [5].

In this paper, we explore HCP's perspectives on self-management mHealth solutions and PGHD as means to support chronic patients' wellbeing, and how these new means are perceived as bringing value to standard care practices in monitoring and treatment.

2. Methods

The study was performed in a context of a larger design research process examining a mHealth solution for self-management of Multiple Sclerosis (MS) [6] called More Stamina. As a part of the solution's development process, a qualitative research approach was taken to conduct structured interviews with HCP from 3 different OECD[2] countries with overall good health care quality indicators: Finland, Spain, and Switzerland. Purposive sampling was used to invite HCP from the Neurology department of the Oulu University Hospital, the Multiple Sclerosis Unit at the Vithas Nisa Sevilla Hospital, and the neurological rehabilitation clinic Kliniken Valens. The main inclusion criterion was that participants should have direct experience working with MS care.

The goal of the More Stamina project is to explore how digital health solutions can survive beyond "laboratory" conditions and be integrated into healthcare systems. The mHealth solution is a gamified task organization tool, designed to help persons with MS manage the impact fatigue has on their daily life. The tool acts as a to-do list where users can input tasks they want to accomplish daily. Besides registering, analyzing, and creating trends for patients´ regular activities, the app tracks activity levels through the in-built sensors, counting steps, walking pace, distances, and GPS positioning. Users control what information to disclose and with whom.

Interviews in Finland and Spain were conducted by native speakers while interviews in Switzerland were conducted in English, with the assistance of a support person that could translate to Swiss German. The HCP were shown a short video featuring the More Stamina solution and later interviewed. Interviews were digitally recorded and translated into English for their analysis. The complete interview guide included questions regarding the overall care pathway of MS; for this paper, we focus only on the responses towards mHealth as means for coping with MS and how mHealth and PGHD could be used to assist HCP in improving care and follow up.

A preliminary analysis of the interviews from Finnish and Spanish HCP was carried out using an Affinity Diagram [7]. The Affinity Diagram is a method for summarizing and structuring large amounts of disorganized verbal data into manageable groups based on natural relationships. This analysis process was carried out by 4 researchers with health and computer science backgrounds. The analysis was performed in 3 workshops sessions. In the first 2 sessions, each interview was reviewed, with main observations and topics raised by the interviewees being extracted; categories were created through emerging topics resulting in an initial categorization. On a third session, topics and categories went through a final revision.

The interviews of the Swiss HCP were conducted on a second stage and were reviewed to identify whether new topics or categories had emerged. It was verified that

[2] Organisation for Economic Co-operation and Development

the observations from the new interviews could be integrated smoothly into the existing categories. The final Affinity Diagram was validated by the whole research team.

3. Results

A total of 23 HCP were interviewed (8 from Finland, 7 from Spain, and 8 from Switzerland). The professions of the interviewees included nurses, physiotherapists, neurologists, neuropsychologists, psychologists to occupational therapists. Clinical researchers and social workers were also included, due to their experience with MS care.

After the analysis, we found that the overwhelming majority of HCP (91.3%) considered that self-management mHealth solutions like More Stamina would be a useful tool to improve their patients' self-awareness, motivation, and overall wellbeing.

More than two-thirds (69.6%) felt that PGHD could assist HCP, whether themselves or other medical profiles, in the role in caring for MS patients (Figure 1).

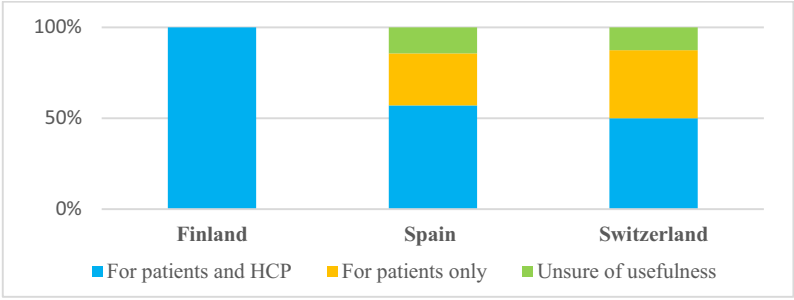


Figure 1. HCP´s views of usefulness of self-management mHealth by country

More than half of HCP (52.2%) showed a personal interest in the possibility of integrating PGHD into care. This group had diverse perceptions of possible uses for PGHD (Figure 2). In general, HCP appeared relatively motivated at the prospect of having access to PGHD that could potentially give them insights into their patients' behaviors or activities. Nevertheless, they appeared less confident in weighing what types of data could be collected through digital solutions, or even how that data could efficiently be presented for easy decision-making. Some of them though did have clearer expectations of the types of information they would like to get from PGHD. Their emphasis laid on:

- Significant changes to the patient's baseline indicators
- Summary of the change in energy levels in a given period
- Fatigue inducing or triggering activities
- Reports on sleep quality, heart rate, and physical activity amount

An 8.7% of HCP reported to be already using PGHD, to some extent, to monitor their patients' health behaviors, as some patients with MS had resorted to their mHealth apps reports to show how much they have walked or slept when asked at consultations

The HCP´s observations on PGHD produced the following subthemes:

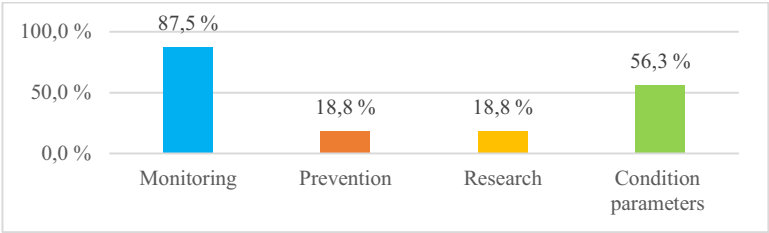


Figure 2. HCP´s Perspective on possible uses for PGHD

3.1 Monitoring and follow-up purposes

Most HCP considered that summaries from PGHD related to physical activity, sleep quality, heart rate variations, medication intake, and other metrics could give them a better understanding of factors affecting their patients in between appointments, especially when there are changes in treatment and medication.

3.2 Prevention of patient deterioration

Some HCP believed that mHealth can provide data that could assist them in the early identification of patterns that sign decline or imminent flare-ups. Prompting patients and HCP with alerts to take action was deemed very appropriate.

3.3 Collection of data for condition research and decision-making

Secondary usage of PGHD was relevant to some HCP for further research on chronic conditions evolution and decision-making in health care workflows and practices.

3.4 Increased transparency of specific condition-specific parameters

Most HCP considered that mHealth solutions as More Stamina could be a useful tool to assess condition-specific parameters, such as fatigue for individual patients with MS. It was suggested that PGHD could be used to help define more measurable indicators of fatigue levels and support in inferring whether patients' fatigue was physical or cognitive.

4. Discussion

The findings in the present study seem to be in line with previous research that confirms HCP views on mHealth solutions as additional support tools for remote management [5] and believe that data provided by digital technologies creates opportunities to better understand factors affecting patients' health outside of the clinical settings, by supporting remote monitoring, expanding care for individuals who have limited access and to improve care for those with acute or chronic conditions [3]. However, an increased use of digital tools seems to evoke concerns in HCP as they feel it might create additional demands and higher workloads in their jobs [8]. Despite those concerns, our findings suggest that HCP in countries with good quality health care indicators appear to have high expectations about the prospect of using digital technologies for patients' self-management in care practices and they seem to be interested in being able to analyze

data that might give them insights into their patient behaviors outside of the clinic. Nonetheless, the big majority of them are not yet using PGHD or do not have precise ideas on how they could use it practice. At least half of HCP appear to view value in the use of PGHD, however, most of them also see a challenge in making it easily understandable and usable.

5. Conclusions

Although perceptions by country varied, on average more than half of HCP considered that self-management mHealth solutions, and the data derived from them, could benefit patients and HCP. They considered that the advantages of using PGHD lay in knowing more about outpatient behaviors and activities, which could help HCP better understand factor affecting patients' health and how treatment could be adapted accordingly. There seems to be positive expectations from HCP about the use of self-management mHealth and PGHD and its possibilities for care. Further research is required to clarify how chronic patients´ PGHD could be better interpreted modeled and presented, using scientific-based approaches, to be efficiently integrated into health care practices in ways that it can make sense to HCP and can bring value to the patient-professional relationship in aspects as monitoring, treatment, and shared decision making.

Acknowledgments

The More Stamina project has received funding from Business Finland.

References

[1] Liao Y, Thompson C, Peterson S, Mandrola J, Beg MS. The Future of Wearable Technologies and Remote Monitoring in Health Care. Am Soc Clin Oncol Educ Book. 2019 Jan;39:115-121. doi: 10.1200/EDBK_238919. Epub 2019 May 17. PMID: 31099626; PMCID: PMC8325475.

[2] Granja C, Janssen W, Johansen MA. Factors Determining the Success and Failure of eHealth Interventions: Systematic Review of the Literature. J Med Internet Res. 2018 May 1;20(5):e10235. doi: 10.2196/10235. PMID: 29716883; PMCID: PMC5954232.

[3] Lavallee DC, Lee JR, Austin E, Bloch R, Lawrence SO, McCall D, Munson SA, Nery-Hurwit MB, Amtmann D. mHealth and patient generated health data: stakeholder perspectives on opportunities and barriers for transforming healthcare. Mhealth. 2020 Jan 5;6:8. doi: 10.21037/mhealth.2019.09.17. PMID: 32190619; PMCID: PMC7063266.

[4] Bardhan I, Chen H, Karahanna E. Connecting systems, data, and people: A multidisciplinary research roadmap for chronic disease management. MIS Quarterly. 2020 Mar;44(1):185-200.

[5] Giunti G, Kool J, Rivera Romero O, Dorronzoro Zubiete E. Exploring the Specific Needs of Persons with Multiple Sclerosis for mHealth Solutions for Physical Activity: Mixed-Methods Study. JMIR Mhealth Uhealth. 2018 Feb 9;6(2):e37. doi: 10.2196/mhealth.8996. PMID: 29426814; PMCID: PMC5889817.

[6] Giunti G, Rivera-Romero O, Kool J, Bansi J, Sevillano JL, Granja-Dominguez A, Izquierdo-Ayuso G, Giunta D. Evaluation of More Stamina, a Mobile App for Fatigue Management in Persons with Multiple Sclerosis: Protocol for a Feasibility, Acceptability, and Usability Study. JMIR Res Protoc. 2020 Aug 4;9(8):e18196. doi: 10.2196/18196. PMID: 32749995; PMCID: PMC7435635.

[7] "Build an Affinity for K-J Method - ProQuest." https://www.proquest.com/docview/214763444?pq-origsite=gscholar&fromopenview=true (accessed Jan. 19, 2022).

[8] Öberg U, et al. Swedish primary healthcare nurses' perceptions of using digital eHealth services in support of patient self-management. Scand J Caring Sci. 2018 Jun;32(2):961-970. doi: 10.1111/scs.12534. Epub 2017 Sep 28. PMID: 28960451.

Challenges of Trustable AI and Added-Value on Health
B. Séroussi et al. (Eds.)

doi:10.3233/SHTI220578

Towards a Computational Approach for the Assessment of Compliance of ALCOA+ Principles in Pharma Industry

Marta DURÁ[a,] and Ángel SÁNCHEZ-GARCÍA[a], Carlos SÁEZ[a], Fátima LEAL[b,c], Adriana E. CHIS[c], Horacio GONZÁLEZ-VÉLEZ[c], and Juan M. GARCÍA-GÓMEZ[a]
[a] *Biomedical Data Science Lab, Instituto Universitario de Tecnologías de la Información y Comunicaciones, Universitat Politècnica de València, Camino de Ver s/n, Valencia 46022, Spain*
[b] *REMIT, Universidade Portucalense, R. Dr. António Bernardino de Almeida 541, 4200-072 Porto, Portugal*
[c] *Cloud Competency Centre, School of Computing, National College of Ireland, Mayor Street, IFSC, Dublin D01 K6W2, Ireland*

Abstract. The pharmaceutical industry is a data-intensive environment and a heavily-regulated sector, where exhaustive audits and inspections are performed to ensure the safety of drugs. In this context, processing and evaluating the data generated in the manufacturing lines is a relevant challenge since it requires compliance with pharma regulations. This work combines data integrity metrics and blockchain technology to evaluate the compliance-degree of ALCOA+ principles among different levels of drug manufacturing data. We propose the DIALCOA tool, a software to assess the compliance-degree for each ALCOA+ principle, based on the assessment of data from manufacturing batch reports and its different levels of information.

Keywords. Data integrity, ALCOA+ compliance, pharma manufacturing industry

1. Introduction

The pharma manufacturing industry is a data-intensive environment that generates large amounts of distributed data regularly accessed by different internal and external stakeholders including international and national regulatory bodies. However, this is hardly a new problem: since the early 1960s when the initial Good Manufacturing Practices (GMPs) [1] for finished pharmaceuticals were published, distinct regulatory bodies have assembled a considerable number of guidelines pertaining to data integrity in pharma manufacturing.

Despite the pharmaceutical industry has consistently improved its manufacturing processes in compliance with good manufacturing practices, it is well documented that falsification of medicines continues[2] and has led to disastrous consequences worldwide[3]. Consequently, different organizations have proposed standards, measures, and protocols to avoid these falsifications. The EU Falsified Medicines Directive[4] introduces harmonized European measures to fight these medicine falsifications and ensure that medicines are safe and that the trade in medicines is rigorously controlled."

Such obligatory safety features, legal framework, and record-keeping requirements have arguably imposed stricter controls for the manufacturing of medicines.

In this context, the gold standard adopted by the pharmaceutical industry is “Data Integrity and Compliance with current Good Manufacturing Practices", defined by the FDA[5], which defines the term “ALCOA+” as a set of principles that should be followed throughout the data life cycle for achieving data integrity. These principles stand that data should be Attributable, Legible, Contemporaneous, Original, and Accurate. Moreover, good documentation practices require that the records are Complete, Consistent, Enduring, and Available.

This work proposes a computational approach for the assessment of these nine ALCOA+ principles among the data generated during the process of drug manufacturing, in order to provide a quantitative measurement of data integrity compliance level. This work has been developed under the Smart Pharmaceutical Manufacturing project (SPuMoNI), a European research project launched by CHIST-ERA pathfinder programme. SPuMoNI consortium includes industry partners, such as a Contract Manufacturing Organisation (CMO), which has been a real scenario for the development of this work.

2. Methods

2.1 Pharma manufacturing reports

At CMOs, the process of manufacturing a particular drug is performed by following the Recipe, which is the protocol that describes in detail the fabrication process of the drug. It is composed of a set of Phases, and each Phase is composed of a set of Instructions. An Instruction is a single action implemented within the manufacturing process. There are various types of Instructions, such as setting a mixing machine or verifying the quantities of raw materials. All this information must be extensively documented and is expected to be ALCOA+ compliant, since regulatory bodies or any other auditor could require an exhaustive revision at any time.

We propose to structure this manufacturing data in what we call a batch Report[6]. At the main level of a Report, the attributes are related to the batch information, such as the batch code, the Recipe code, and the Qualified Person; which is responsible for assuring the quality of the manufactured drugs that will be available on the market. Furthermore, a Report contains the list of materials used for the process as well as the row data produced in the production line. This information is organised in the Report following the Recipe structure: a) a list of Phases that contains a set of Instructions; b) each Instruction item includes a list of parameters to be controlled during the execution and the data recorded during the process (*e.g.*, temperatures or mixing speeds).

2.2 ALCOA+ principles assessment

Following the definitions of ALCOA+ and data integrity methods[6], we have defined a set of metrics (Table 1) for evaluating the compliance of each principle in a batch Report. These metrics are implemented in the tool proposed in this work (Results Section) that we have named “DIALCOA” (**D**ata **I**ntegrity **ALCOA** assessment).

Table 1. ALCOA+ Principles definition and their proposed metrics for assessing its compliance in batch manufacturing Reports

	ALCOA+ Principle	Proposed Metric
Attributable	All data must be attributable to the person generating the data including who performed an action and when.	Assessed by measuring the amount of data which have been assigned to the person who did the collection and the identification of the person responsible of the report Attributable score is the percentage of data which has been recorded by the staff who has collected it.
Legible	All data must be legible and permanent. Ensuring records are legible	Assessed among all Report fields by the quantification of measurements that comply with these three specifications: - Data must be electronic and use UTF-8 format - Decimal numbers must use the same format - Free texts must use words present in language dictionaries Legible score is the percentage of data that is compliant with the three specifications.
Contemporaneous	Record the data at the time it is performed. Date and time stamps should flow in Date and time stamps should flow considering the execution order to prove the data credibility.	Assessed by verifying and counting the Report fields which include the date and time of data creation. Contemporaneous score is the percentage of data that includes its timestamp.
Original	The preservation of original records, to verify the authenticity of the data.	Assessed by verifying that the Report has not been adulterated. To do so, DIALCOA relies on a blockchain Smart Contract, where the original version of the Report is stored. Original score is 100 if data is original or 0 if any field of the Report has been adulterated.
Accurate	Data should be free from errors. Editing should only be performed by using the principles of GDPs.	Assessed by range checks and outlier detection methods. Those numerical fields of the Report are checked with the expected range of acceptable values in the Recipe, and to detect outlying behaviour. Accurate score is the percentage of numerical fields of the Report that satisfy the rules above.
Complete	All data must be retained from the creation of the documentation. Deletion or removal of data must not take place.	Assessed by checking that all expected fields in the Report are fulfilled. Complete score is the percentage of fulfilled fields among the Report.
Consistent	Data must be coherent with the expected information in time.	Consistency principle can be assessed by evaluating time consistency, counting the tracking start date that is earlier than tracking end date for all the Report. Consistency score is the percentage of data that is compliant with this time consistency rule.
Enduring	Store systems are running during all life of data.	Assessed by requesting a certified expiration date of the Report. Enduring score will be 100 if the expiration date is included and updated and 0 if it is missing or if it has expired.

Available	Data must be accessible at any moment it is request.	Assessed by requesting a certified expiration date for accessing to the Report. Available score will be 100 if the expiration date is included and updated and 0 if it is missing or if it has expired.

2.3 Blockchain private network

The DIALCOA tool is connected to a blockchain private network that is composed of a private Ethereum network infrastructure[7]. When a Report is uploaded to DIALCOA, an originality assessment is performed by uploading a new batch record to the Ethereum network as a smart contract and verifying the uniqueness of all its data. Based on the previously uploaded reports on the Ethereum network, the originality score is calculated evaluating the uniqueness of the new data by comparing it with the existing stored information.

3. Results

This work presents the first steps towards a computational approach to assess the compliance of ALCOA+ principles within batch manufacturing data. This software implements the proposed metrics described in Table 1 to be used by the Qualified Person at the CMO to monitor the integrity of the data that have been generated.

This software can be installed in the pharma manufacturing plant systems to access production data and batch Reports. Additionally, a private blockchain network should be installed in order to ensure the traceability of the ALCOA+ assessments and the Original principle evaluations. Figure 1 outlines the information workflow and the connection among elements.

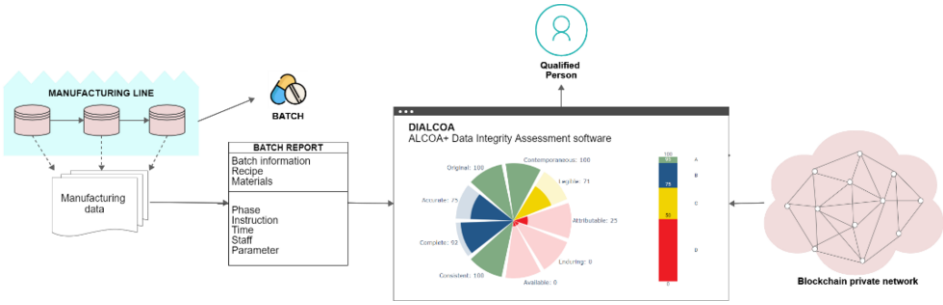


Figure 1. Workflow of DIALCOA tool in pharma manufacturing process

DIALCOA shows a global view of the nine ALCOA+ principles scores in a pie chart, including a color scale for the limits of compliance (Figure 1). Moreover, the user can explore the detailed analysis of each principle assessment. This is possible since the software is able to detect and plot the potential data integrity conflicts which are causing scores lower than 100. Hence, the user can easily identify which Report data present data integrity issues for each ALCOA+ principle.

Discussion

The proposed system feasibly supports the compliance of ALCOA+ principles by evaluating batch Reports through data integrity metrics. To achieve a higher readiness level, an evaluation of the proposed tool in the pharma shop-floor environment is being performed. As future work, we aim to validate DIALCOA tool in a real pharma manufacturing environment.

4. Conclusion

The pharmaceutical industry is a data-intensive and heavily regulated domain. Its manufacturing lines continuously generate large amounts of data that must be collected and have to be ALCOA+ compliant. This industry requires effective solutions to improve its manufacturing process in terms of data integrity compliance. Towards this scenario, we propose a novel tool for assessing the compliance of ALCOA+ principles within batch manufacturing reports.

Acknowledges

We thank the consortium of Smart Pharmaceutical Manufacturing project (https://www.spumoni.eu/) for enabling the development of this work in a real scenario of pharma industry. We also thank the co-funding support of European Union and Spanish Agencia Estatal de Investigación (PCI2019-103783) for the development of SPuMoNI project.

References

[1] US Federal Register–Part 133, 28 FR 6385, 20/Jun/63. Drugs, Current Good Manufacturing Practice in Manufacture, Processing, Packing or Holding. American Journal of Hospital Pharmacy. 1964 09;21(9):398–401

[2] D. McManus, B. D. Naughton, A systematic review of substandard, falsified, unlicensed and unregistered medicine sampling studies: a focus on context, prevalence, and quality, BMJ Global Health 5 (8) (2020) e002393. doi:10.1136/bmjgh-2020-002393.

[3] M. S. Rahman, N. Yoshida, H. Tsuboi, N. Tomizu, et al., The health consequences of falsified medicines-a study of the published literature, Tropical Medicine & International Health 23 (12) (2018) 1294–1303. doi:10.1111/tmi.13161.485

[4] European Parliament–Council of the European Union, Directive 2011/62/EU of the European Parliament and of the Council of 8 June 2011 amending Directive 2001/83/EC on the Community code relating to medicinal products for human use, as regards the prevention of the entry into the legal supply chain of falsified medicinal products, Official Journal of the European Union 54 (2011) Document 32011L0062. doi:10.3000/17252555.L_2011.174.eng

[5] US Department of Health and Human Services. Data Integrity and Compliance With CGMP Guidance for Industry. Silver Spring: Food and Drug Administration; 2016. Pharmaceutical Quality/Manufacturing Standards (CGMP)

[6] Leal F, Chis AE, Caton S, González-Vélez H, García-Gómez JM, Durá M, et al. Smart Pharmaceutical Manufacturing: Ensuring End-to-End Traceability and Data Integrity in Medicine Production. Big Data Research. 2021;24:100172. (Open Access).

[7] Leal F, Chis AE, González-Vélez H. Performance Evaluation of Private Ethereum Networks. SN Computer Science. 2020;1(5):285

Challenges of Trustable AI and Added-Value on Health
B. Séroussi et al. (Eds.)

doi:10.3233/SHTI220579

Are People Ready to Report Digital Health Ethical Issues in Order to Contribute to Their Resolution?

Magali ROBERT[a], Nathalie BAUDINIERE[a] and Brigitte SEROUSSI[b,c,1]
[a] *Agence du Numérique en Santé, Paris, France*
[b] *Sorbonne Université, Université Sorbonne Paris Nord, INSERM, Laboratoire d'Informatique Médicale et d'Ingénierie des connaissances en e-Santé, LIMICS, F-75006 Paris, France*
[c] *AP-HP, Hôpital Tenon, Paris, France*

Abstract. Although guaranteed by the GDPR, transparency of health data processing may not be fully respected, leading citizens to mistrust eHealth and discard digital health services. Identifying and safeguarding ethics in eHealth services is thus important to promote their development. We conducted a survey to assess the extent of ethical issues induced by the use of digital health services, understand the efforts citizens would be willing to accept for reporting such issues, and evaluate citizens' expectations regarding this reporting. Among 200 respondents, 36% reported having encountered ethical issues with the processing of their health data or with digital health services being poorly inclusive. Faced to ethical issues when using a digital health service, 49% of respondents were rather or very angry, and 33% felt rather or very dependent. Most respondents were ready to report digital health ethical issues if there is a feedback for each report.

Keywords. eHealth services, ethics, health data processing, online survey

1. Introduction

Digital technologies have been introduced, and sometimes imposed, at great speed in many areas of the daily lives of French people, particularly in the field of health. If digital health technology does promise benefits (*e.g.*, helping people book medical appointments more efficiently, completing administrative procedures for pre-admission in a healthcare facility, enable remote consultations, or being monitored continuously outside of consultations), eHealth is not exempt from general concerns that this accelerated transformation could deepen inequalities [1]. In addition, several ethical concerns are associated with the role of big data analysis in personalized healthcare [2]. Identifying and safeguarding ethics in eHealth services is thus important to promote the development of an ethical eHealth [3], because it is also not ethical to not offer eHealth services benefits to all.

Beneficence, non-maleficence, justice, and autonomy are important overall ethical principles enshrined in the Hippocratic Oath. When crossed with digital health technology, these ethical principes essentially cover issues related to privacy, protection,

[1] Corresponding author, Brigitte Séroussi; E-mail: brigitte.seroussi@aphp.fr

defence of rights, advocation, equality and equity, and trust [4]. Trust relies on data processing transparency (citizens should find out who accessed their health data, when and for which purpose) that may reassure citizens about their data being processed properly. However, though guaranteed by the GDPR, these rights may not be fully respected leading citizens to mistrust eHealth and discard digital health services.

Prior to the launching of a new national French digital health service, My Health Space, a digital health record system accessible to patients and authorized healthcare professionals, and the expected concomitant deployment of digital uses in health, the French Ministry of health wanted to assess the interest of setting up a national platform for reporting digital health ethical issues. The objectives of this work are to assess the extent of ethical issues arising by the use of digital health services, identify the populations that are most likely to experience these issues (patients? health professionals?), understand the efforts that citizens would be willing to accept to report an ethical issue, and evaluate the expectations of citizens following this reporting.

2. Material and Methods

2.1. Conclusions of prior national citizen workshops

In 2020, the French Ministerial Delegation for Digital Health organized citizen workshops that concluded on four major areas of concern about digital health: (i) exclusion of a portion of the population (digital divide), (ii) fragmentation of health information (the silos of a fragmented health system lead to a lack of fluidity in an often hyperspecialized medicine, which involves a large number of medical and administrative actors in different temporalities, (iii) deprivation of patient information and lack of transparency of data processing, (iv) complexity of digital health services.

2.2. Analysis of citizens' need of a tool to report digital health ethical issues

We first organized face-to-face semi-structured interviews of a sample of digital health users, including citizens and healthcare professionals. The interview script addressed questions on the participant's actual use of digital health services, the ethical issues he/she encountered, and how he/she reacted (did he/she try to find solutions on internet? did he/she try to report the issue (e.g., through customer service), what reporting services did he/she use?). Each interview was recorded in a report. All reports were transcripted for manual thematic analysis [5].

In a second step, we conducted an online anonymous survey targeting patient-citizens experiencing a daily use of digital health tools in the follow-up of their chronic disease. The questionnaire was published on social media, relayed by *France Asso Santé* (the French National Union of registered associations of users of the health system). The survey asked about (i) digital health services most frequently used (prevention and wellness tools such as calorie tracking, step count calculation, pulse tracking; websites to consult and download laboratory or medical imaging results; online appointment scheduling; online pre-admission/file submission for medical or medico-social clinical facilities; monitoring of chronic diseases, e.g., diabetes, asthma, or of drug side effects, e.g. oral chemotherapy; monitoring of biological constants as part of a medical follow-up, e.g., blood sugar, blood pressure; personal medical record, such as the French DMP, or EHRs), (ii) ethical issues encountered by citizens concerning their health data (I didn't

understand where my health data were going; I didn't understand if my health data were shared and with whom; I didn't understand if my health data were reused to be sold by the digital solution provider; I tried to delete information but could not; I was asked for personal information that was not relevant to my condition or the purpose of the digital health service), (iii) technical difficulties experienced by citizens when using a digital service (I could not find my information in the application; I couldn't use the tool or the website because it was too complicated; I read the terms and conditions (T&C), but I did not understand what kind of data processing was performed and whether data were shared with third parties), (iv) feelings of citizens when faced to digital health ethical issues (Lonely; Stressed/worried; Dependent; Upset, all items being graded on a 5-point Likert scale, with 1: not at all, and 5: very), (v) citizen reactions (I did nothing, but I did not use the digital solution anymore; I did nothing and deleted the application; I talked about the issue with a health care professional; I searched the internet to find where I could describe my problem; I reported the issue on social media (forums) to see if other people encountered the same issue and how they solved it).

The last question was to test if there was a portal (developed by the government) where one could report digital health ethical issues with the aim of contributing to their resolution, one would use it or not and why.

3. Results

3.1. Face to face interviews

Individual interviews were conducted between September and October 2021 on a sample of 32 people selected to represent the diversity of digital health users and that accepted to participate: health professionals, citizens, experts from medico-social and health sectors (including a director of a long-term care facility), someone from a healthcare technology company (Doctolib), from CNIL (*Commission nationale de l'informatique et des libertés* is the French regulator of personal data), representatives of patient associations (France Asso Santé and the French Diabetes Federation), and representatives of hospital patients. Each interview lasted between 30 and 90 minutes.

Digital health services mostly used were mainly online appointment booking and laboratory results retrieval sites, with a high level of confidence in these sites. Citizens did not raise any ethical issues. Despite they agree that how things are working is "opaque", their suspicion was compensated by their trust towards health professionals (for the management of EHRs) and the value of the service provided (health or well-being websites and applications). The most confident were those who work in the digital sector, because they felt that they have a better understanding and control of the process. Fears were finally distant threats/disagreements as compared to the service provided and the health issues raised by digital tools.

3.2. Online survey

Between October 22, 2021 and November 5, 2021, 400 responses were collected, but we excluded the first 200 from our statistics because of some inconsistencies in the questionnaire. In both the interviews and the online survey, the majority trend was the use of online appointment booking and laboratory results retrieval sites. Figure 1 displays the distribution of mostly used digital health services.

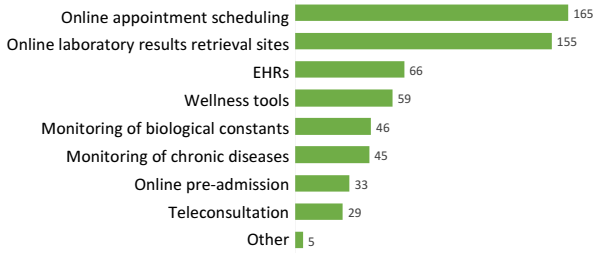


Figure 1. Distribution of mostly used digital health services (n=200).

Of the 200 actual respondents, 63 reported having encountered ethical issues (on health data processing or other topics such as the weak inclusivity of digital health services). Figure 2 displays the distribution of such ethical issues.

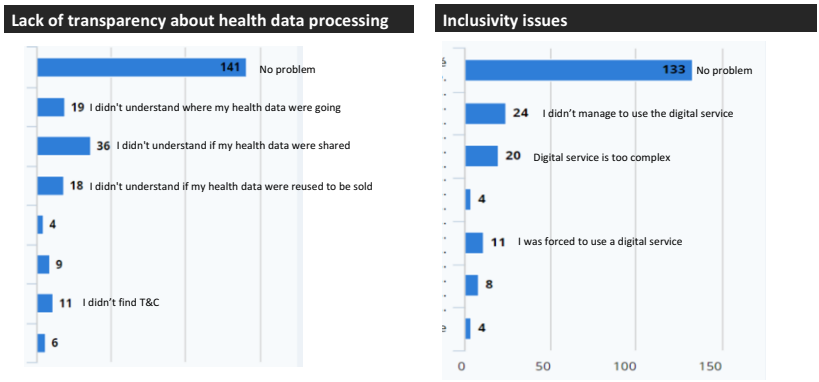


Figure 2. Distribution of ethical issues for health data processing and digital health accessibility (n=200).

Other digital health ethical issues have been reported in the "other" answer:

- Appointment scheduling for people with disabilities: "When I arrived at the appointment in a wheelchair, I had to go back home because I had not specified that I was in a wheelchair, which was impossible when I made the appointment online, and there was no access for disabled persons."
- Forced teleconsultation: among teleconsultation users, more than half complained that they were forced to use the service (n=29).
- Access to online laboratory results: "I gave up on getting my exam reports because of problems accessing the service."
- Forced consultation: "The replacement of my physician called me even though I am not followed for a particular pathology. I discovered afterwards hat he billed for a teleconsultation, even though he had not informed me.

The analysis of the answers to how respondents felt when faced to some ethical issues when using a digital health service showed that 49% of the persons were rather or very angry, and that 33% felt rather or very dependent (see Figure 3).

The survey provides answers to how people react when trying to solve ethical issues. Of the people who said they encountered one of the problems listed, 27% (out of 63)

took advantage of the online questionnaire to describe it in the "Specify" window. When questioned about whether they would report digital health ethical issues on a dedicated governmental platform, some respondents give conditions and obstacles: "have a feedback", "reports need to be analyzed", "one more reporting service!"

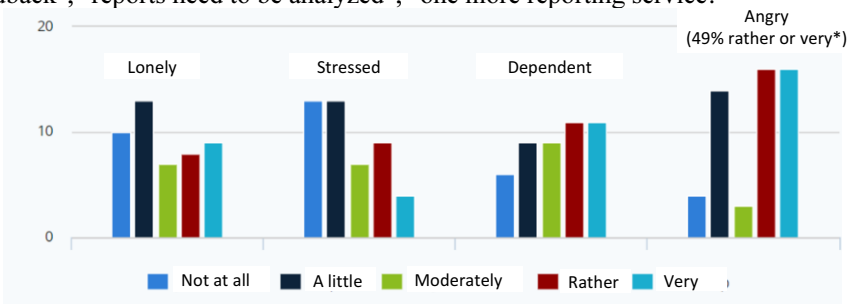


Figure 3. Distribution of users' feelings when faced to digital health ethical issues (n=63)

4. Discussion and Conclusions

Users encounter known digital health ethical problems, but the details for action have yet to be defined. Existing media (*e.g.*, CNIL complaint service) allow citizens to report irritants, but they are little or not treated. A significant part of digital health users was nevertheless ready to report the problems encountered, only if some conditions are checked: there is a feedback for each report; the ethics unit in charge of processing the reports should strongly communicate to encourage compliance with the laws and regulations in force by the actors concerned (professionals, publishers, organization); the reporting service must be visible and easily accessible from the tools concerned: it is necessary to find a high-traffic site that makes the service visible and effortlessly accessible; the implementation of a way of processing reports according to their types is essential for a quality user service which requires a significant work to identify and characterize these types; having a tool for processing reports must be anticipated before getting high volumes (CRM type or other).

References

[1] Jokinen A, Stolt M, Suhonen R. Ethical issues related to eHealth: An integrative review. Nurs Ethics. 2021 Mar;28(2):253-271.
[2] Althobaiti K. Surveillance in Next-Generation Personalized Healthcare: Science and Ethics of Data Analytics in Healthcare. New Bioeth. 2021 Dec;27(4):295-319.
[3] Skär L and Söderberg S. The importance of ethical aspects when implementing eHealth services in healthcare: A discussion paper. J Adv Nurs. 2018 May;74(5):1043-1050.
[4] De Rosis S and Barsanti, S. Patient satisfaction, E-health and the evolution of the patient-general practitioner relationship: evidence from an Italian survey. Health Policy 2016; 120(11): 1279–1292.
[5] Braun V and Clarke V. Using thematic analysis in psychology, Qualitative Research in Psychology 2006,3(2):77-101.

Challenges of Trustable AI and Added-Value on Health
B. Séroussi et al. (Eds.)

doi:10.3233/SHTI220580

Using Robots in Medical Informatics Education

Thomas BÜRKLE[a,1], Matthias BENDER[a] and Robin GLAUSER[a]
[a]*Bern University of Applied Sciences, Institute for Medical Informatics*

Abstract. Although robots have been used for quite some time in education on school and university level, we found no reports of robots being used in the teaching of medical informatics. Thus we present the timetable and structure of a one week, 2 ECTS blocked course for robots in medical informatics initiated in autumn 2022. 19 participants completed the first iteration. We report about the requirements in terms of an appropriate programming environment, the combination among robots and our medical informatics lab and the results of the student's evaluation of the first instance as well as the experiences with the two types of robots used.

Keywords. Robots in medical informatics, teaching

1. Introduction

Robots have been used for some time in education, often starting at school level and commercial products such as NAO are available [1,2,3]. In a scoping Medline search no reports of robots used in the medical or medical informatics education have been found, although the use of robots in medicine has been reported e.g. for surgery [4], neuro-rehabilitation [5], cerebral palsy [6], and as learning systems for improved health literacy [7]. In the ongoing SARS-CoV-2 pandemic, Cruzr robots have been used as a first line to detect the body temperature of visitors or patients and interact with them in multiple languages [8, 9].

Bern University of Applied Sciences (BFH) offers a BSc in Medical Informatics since 2011 [10]. The three-year program comprises 180 ECTS and has recently advanced to a new study plan. Furthermore, the faculty and institute of medical informatics provide a medical informatics lab comprising nearly all institutions of the Swiss healthcare system such as inpatient care including surgical theatre and intensive care, GP practice, pharmacy and an AAL environment for research and teaching purposes. A novel hands-on course using robots has been introduced into the Medical Informatics study.

2. Methods

The so-called block-week is a new format valued 2 ECTS within the remodeled BSc study. Students will complete three block-weeks at different levels. The goals for the block-weeks are

[1] Thomas Bürkle, Bern University of Applied Sciences, Quellgasse 21, 2501 Biel/Bienne, Switzerland; E-mail: thomas.buerkle@bfh.ch.

- To perform interdisciplinary projects
- To select adequate methods from the different subject areas for realization
- To organize teamwork

From 15th to 19th Nov 2021 the second year students (n=19) had for the first time a block-week "robots in medicine". Two Cruzr [11] Version 03H18001/3.7.6 and two Pepper [12] version 1.8a with Python API on NAOqui 2.5 have been provided for four student groups with a group size four to five. Three instructors supported the course. Fig 1 is taken in the course.

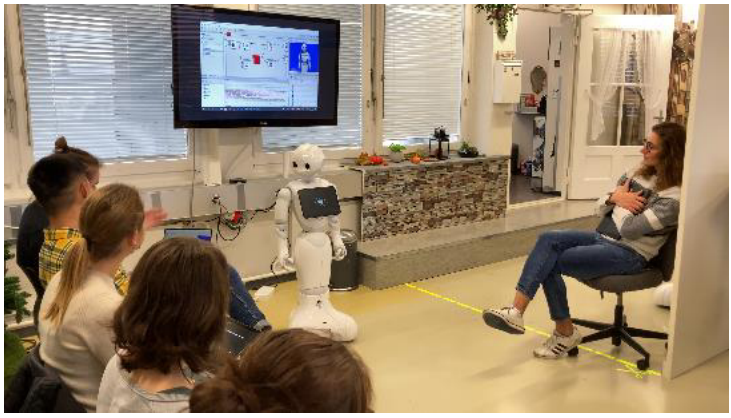


Figure 1. A group of students in introduction to visual Pepper programming interface

The following assignment was given in 2022: "Establish a medicine-related guided visitor tour in the BFH medical informatics lab". A competition was announced for the best lab tour to win a prize. The task was split in daily subtasks. Monday, the student groups should be able to move the robot in the lab and let it tell first sentences. Tuesday they should be able to move each joint of the robot, to let the robot navigate throughout the lab and to make it touch a mechanical door switch. Wednesday was dedicated to basic speech recognition, so that the robot would interact with the user. Thursday was reserved for freestyle programming. Friday morning was a joint activity of block-week #2 (robots) and block-week #3, who had performed a tour de Suisse to various hospitals and Medical informatics suppliers, and comprised the competition for the best lab tour with a timeslot of 15 minutes for each group.

Each day, accompanying plenaries were organized with external speakers, e.g. a) how does the Swiss company raumcode employ robots, b) which Swiss pharmacy data resources and services are available, c) how could the Laerdal patient simulator interact with the robot. The course timetable is provided in fig 2.

A generic control interface over websockets named RobotControlAPI was implemented by RG to enable the students for "detached" layered programming of the Cruzr robots. A client implementation for Python was developed as well for the students to use. With this client the students could develop code in Python which triggers functions of the Ubtech Cruzr. The Ubtech Cruzr is split into two parts, a tablet with Android 5.0 for the user interface on which the websocket interface was programmed and a ROS deployment on Ubuntu 18.04 for the motor functions and navigation. Program

Code of all groups was collected in GIT. For the Pepper robots the Choregraph 2.5.10 programming interface was recommended, but Python 2.7 SDK was enabled as well.

	Monday 15th	Tuesday 16th	Wednesday 17th	Thursday 18th	Friday 19th
08.30-10.00	Intro Demonstrate Cruzr labtour group-building plan for the week	presentation Mr. Stauffer, raumcode	presentation Mr. Sonnenschein HCI	presentation Laerdal patient simulator	presentations of block-week #3 visits in hospitals and companies
10.30-12.00	presentation: How to navigate robot	presentation: How to move arm and hand	presentation: language recognition	presentation advanced NLP	competition 4 groups block-weel #2
break					
13.00-17.00	free programming goal: robot talks and moves around	free programming goal: robot opens door and can navigate complete lab, tells story for every station	free programming goal: robot recognizes drug and tells at least dosage and use, communicates with HUE	free programming frestyle competition	prepare final documents and finalize work

Figure 2. Timetable of the block-week #2 in 2022.

3. Results

The final results of block-week #2 and #3 were demonstrated Friday morning. Due to lab space restrictions and COVID restrictions, remote transmission of Video to a large classroom was organized. The best guided lab tour was voted from all students of both block-weeks and a prize was distributed to the members of the successful groups. Two groups using the Cruzr robot presented the full and comprehensive lab tour including individual add-ons. The winning group could show the potential of robot remote activity avoiding the risk of infections. They made their Cruzr ask the tour participants for a valid COVID certificate and check it out before the lab tour started. If the certificate was nonexistent or invalid, the robot would instruct the visitor how to get a COVID test nearby. One Pepper group demonstrated the "Face-tracker" mechanism of the Pepper which follows a person through the lab.

The BFH courses are evaluated on a regular basis with standardized questionnaires. For block-week #2 the evaluation had a good response (13 from 19 = 68%), some of the results are given in table 1.

Table 1. partial evaluation results of the block-week #2, 13 participants, mean values, range 1-4

	Mean on 4 pt Likert scale
I could work independently	1.5
I achieved the set goals mostly	1.5
The instructors supported with constructive critique	1.5
The instructors were open for questions	1.2
Study goals and assessment criteria were timely known	1.8
External presentations delivered new input	2.1
Overall impression of the course	1.6

Results are measured on Likert scale between 1 = best and 4 = worst where 2.5 would be the median. The overall evaluation of block-week #2 was best among all three

block-week events. Individual feedback comprised comments such as “learned a lot” or “activity in the lab was good” or “good working environment”. Some critique came for the Pepper robot hardware and the organization for the two Pepper groups, which could be improved.

4. Discussion

The primary goal of the course was to motivate students to find out how robots could be potentially used in a medical environment. Thus, the competition character, and the winning group demonstrated a nice example mimicking a case where robots might be the first contact of a potential infectious person and perform some kind of triage until the person’s status is confirmed. It was known in advance that the available hardware has limitations and may not be optimal for medical use cases. Both robots are unable to manipulate heavy objects with their hands, in part due to the fact, that for security reasons the finger joints do not give resistance. Both robots do not have a tray and are not well suited to carry heavy loads. In the case of Pepper, load bearing is also hindered by the threat of instability.

Activities such as opening doors are not possible due to the safety design and control options of the extremities of both Cruzr and Pepper. It was disappointing however to see, that even the stripped down task of pushing an electric door opener switch could not be accomplished. It was impossible to move the robot repeatedly in exactly the same position by programming methods. This task was therefore skipped and successfully replaced by WLAN based interaction with HUE lamps and switches, e.g. to turn on the monitor of the radiology viewing station the moment the robot stopped there and explained this application.

Furthermore, the two kinds of robots are different by nature. Pepper is more a “social robot” with an emphasis on verbal and optical interaction with a human present in the room, while Cruzr, derived from industrial cleaning robots, is a good navigation robot supplied with LIDAR sensors. Cruzr can be fed with external map data of the environment and is therefore somewhat better suited towards e.g. leading a lab tour over some distance. One of the Pepper groups circumnavigated the weakness of Pepper in long-distance navigation using the “Face tracker” mechanism. Thus, the visitor would stroll himself through the lab and Pepper would follow him and give comments where appropriate. The other Pepper group limited themselves to the pharmacy environment without moving large distances.

For all groups it was easy to let the robot talk predefined sentences, e.g. explaining the different parts of the lab in intermediate stops. A problem was the pitch and language in the Pepper robots and Pepper groups did not change to German as the conversation language. This was partially improved when the speed of talking was adjusted programmatically. All groups were able to program their robots to understand simple sentences based on catalogued phrases and search words. There is room for improved NLP processing in future iterations of the course using either internal or external AI methods for real-time “intelligent” interaction.

Originally it was planned that the robot should take an arbitrary drug package from the table, scan the barcode data, and, based upon a Swiss pharmacy database, tell the user what the drug was good for and how he should take it. Due to the mentioned problems with physical interaction the task was slightly altered. All groups were able to position a drug package in front of the robot’s built in camera and to recognize the drug.

5. Conclusion

According to the module description, students of block-week #2 should

- Understand through own experience the connection between knowledge and its practical application
- Understand and use their knowledge in practice related tasks and choose the adequate methods from different subject areas for realization
- Judge and solve problems in shape of a practice oriented task which utilizes learned methods

It seems fair to state that these goals could be achieved. Furthermore, evaluation results suggest that the participants enjoyed the format and the given opportunities for independent group work. They learned that programming a robot is not a miracle and that visible effects can be achieved in a short time span. They (and the instructors) gained practical experience with the limitations of current hardware and should now be able to describe the requirements for "medical grade" robots. They experienced the need for fault tolerance in medical programs, e.g. when the robot stopped completely, just because it could not interpret the coded data of a given drug package.

The restricted physical capabilities of our hardware prompted us to acquire a new robot named Lio [13] which is better suited for physical interaction e.g. with drug packages. It will be interesting to see if we need either a variety of specialized robots for each task or if the dream of an universal robot becomes true which can be used for various activities in a medical environment.

References

[1] Rouhianen L. How to use robots in education. Available under https://www.lasserouhiainen.com/how-to-use-robots-in-education/ last visited Mar 14th, 2022

[2] Purdue University. The Use of Robotics and Simulators in the Education Environment. Available under https://online.purdue.edu/blog/education/robotics-simulators-education-environment last visited Mar 14th, 2022.

[3] Inamdar T. Will Robot Teachers be the Future of Education. Available under https://www.linkedin.com/pulse/robot-teachers-future-education-tanveer-inamdar-/?src=aff-lilpar&veh=aff_src.aff-lilpar_c.partners_pkw.123201_plc.adgoal%20GmbH_pcrid.449670_learning&trk=aff_src.aff-lilpar_c.partners_pkw.123201_plc.adgoal%20GmbH_pcrid.449670_learning&clickid=XNcU0Ez4HxyIWrPQHcWEzT3UUkG1PBVylU%3A40c0&mcid=6851962469594763264&irgwc=1 last visited Mar 14th, 2022.

[4] Ferrigno G, Baroni G, Casolo F, De Momi E, Gini G, Matteucci et al. Medical robotics. IEEE Pulse. 2011 May-Jun;2(3):55-61.

[5] Krebs HI, Volpe BT, Williams D, Celestino J, Charles SK, Lynch D, et al. Robot-aided neurorehabilitation: a robot for wrist rehabilitation. IEEE Trans Neural Syst Rehabil Eng. 2007 Sep;15(3):327-35.

[6] Meyer-Heim A, van Hedel HJ. Robot-assisted and computer-enhanced therapies for children with cerebral palsy: current state and clinical implementation. Semin Pediatr Neurol. 2013 Jun;20(2):139-45.

[7] Wei CW, Kao HY, Wu WH, Chen CY, Fu HP. The Influence of Robot-Assisted Learning System on Health Literacy and Learning Perception. Int J Environ Res Public Health. 2021 Oct 21;18(21):11053.

[8] Sarker S, Jamal L, Ahmed SF, Irtisam N. Robotics and artificial intelligence in healthcare during COVID-19 pandemic: A systematic review. Robotics and Autonomous Systems 2021 Dec;146:8.

[9] Lafranca L. and Li J. International Conference on Social Robotics, Springer (2020):688-07

[10] Holm J, Bürkle T, Gasenzer R, von Kaenel F, Nüssli S, Bignens S, et al. A Novel Approach to Teach Medical Informatics. Stud Health Technol Inform. 2015;216:1011.

[11] https://www.ubtrobot.com/de/products/cruzr?ls=en last visited Mar 14th, 2022.

[12] https://www.softbankrobotics.com/emea/de/pepper last visited Mar 14th, 2022.

[13] https://www.fp-robotics.com/de/care-lio/ last visited Mar 14th, 2022.

Challenges of Trustable AI and Added-Value on Health
B. Séroussi et al. (Eds.)

doi:10.3233/SHTI220581

ECHO: An Information System for the Monitoring and Evaluation of Dental Student Activity in a Pre-Doctoral Clinic

Benoit BALLESTER [ac,1], Frédéric BUKIET [ab], Jean-Charles DUFOUR [cd]
[a] *Pôle d'Odontologie, Assistance Publique des Hôpitaux de Marseille, France*
[b] *Aix Marseille Univ, CNRS, ISM, Inst Movement Sci, Marseille, France*
[c] *Aix Marseille Univ, INSERM, IRD, SESSTIM, Sciences Economiques & Sociales de la Santé & Traitement de l'Information Médicale, ISSPAM, Marseille, France.*
[d] *APHM, Hôpital de la Timone, Service Biostatistique et Technologies de l'Information et de la Communication, Marseille, France*

Abstract. The main goal of dental education is to develop health professionals who will maintain and improve the oral health of patients. This requires the quantitative and qualitative assessment of dental student activity. The aim of this work is to describe the ECHO information system that was developed at Timone Hospital (France) for the monitoring and evaluation of dental student activity and to present the results of a qualitative evaluation of student perceptions of this system. According to the analysis of the UML model of care-related procedures and data, the pre-existing process of student evaluation was characterized by redundancy between administrative and educational data. ECHO was developed in PHP/MySQL and designed to centralize the two types of data in a unified computerized process. The qualitative evaluation of dental student perceptions of ECHO was performed using an anonymous online Google Form questionnaire. Among the respondents (102/254 students), 96% stated that ECHO is easy to use, 86% that it saves time, and 81% that it gives them a better overview of their activity. After several years of use, ECHO has solved many of the difficulties related to the use of internship paper booklets, while also providing a documentary database of the activities of our dental department. The student activity data stored in ECHO are directly accessible by faculty members and can be reused to facilitate departmental management and research and to improve patient follow-up.

Keywords. Dental Informatics, Information System, Dentistry, Education.

1. Introduction

The main goal of dental education is to develop health professionals who will maintain and improve the oral health of patients [1]. This implies turning undergraduate students into practicing dentists throughout their preclinical training [2,3], which in turn requires the quantitative and qualitative assessment of their clinical activity. In French pre-doctoral dental clinics, dental care is mainly delivered by undergraduate students under the supervision of licensed practitioners/teachers [2,4]. Depending on the type of care,

[1] Corresponding Author, Ballester Benoit, 27 Boulevard Jean Moulin, 13385 Marseille, France; E-mail: ballester.benoit.1@univ-amu.fr

the acts performed by students are validated by a specific practitioner/teacher. The latter must check the quality of each step of the clinical procedure before the student can proceed to the next step, and must conduct an overall evaluation at the end of the care session. In most such clinics, the clinical activity of each student is recorded in an internship paper booklet that is evaluated in its entirety on a bi-annual basis. This paper booklet, not linked to the hospital information system, make the monitoring of student activity time consuming and represents both an organizational and human resource challenge for faculty members. For students, the time spent on managing and tracking their own activity in the booklet comes at the expense of time spent on clinical learning.

At the Dental Department of Timone Hospital (AP-HM – Assistance Publique Hôpitaux de Marseille, France), a pre-doctoral clinic where about 250 students manage 25,000 patient visits each year, an information system for the monitoring and evaluation of dental student activity was developed to simplify this evaluation process. This system called ECHO was launched in 2018 and has since been adopted by all students and faculty. The aim of this work is to describe the functioning of ECHO and to present the results of a qualitative evaluation of dental student perceptions of this information system.

2. Materials and Methods

ECHO was developed following the analysis of a Unified Modeling Language (UML) model of the care-related (medical, administrative, and educational) procedures and data recorded as part of the pre-existing processes at the end of each care session. This new information system was designed to meet the following requirements: 1) to facilitate the quantitative and qualitative recording of all student activities; 2) to help monitor the completion of mandatory procedures; and 3) to allow for the validation of students' essential clinical skills. It also had to meet the requirements classically expected of computer registries, namely: a) to ensure short training time with easily navigable screens; b) to be web-based; c) to include drop-down menus and logic checks; d) to comply with General Data Protection Regulation and internet security standards; h) to allow for printing out summary data; i) to be linked to but not act as a substitute for electronic dental records (EDRs) [5].

ECHO was registered with the Data Protection Officer (formerly named CIL – *Commission Informatique et Liberté*) of Timone Hospital under #2018-01. Its back-end was developed in PHP/MySQL with the open source MVC framework Laravel [6], and its front-end is based on the opensource framework Bootstrap [7]. The database was designed using Laravel schema designer [8], which allows automatic generation of models and migrations tables for the Laravel framework. ECHO is hosted on an Apache server managed by the digital services department of Timone Hospital.

In the months following the launch of ECHO, a qualitative evaluation of dental student perceptions of this information system was conducted using an anonymous online Google Form questionnaire. The questionnaire was available from 2018/03/13 to 2018/03/27. Students were invited to participate by email, with a recall notice at 7 days.

3. Results

According to the analysis of the UML model of care-related procedures and data, the pre-existing process of student evaluation was characterized by redundancy between

administrative and educational data. Indeed, each type of data was collected separately on different media (i.e., administrative data on paper forms and academic data in the internship booklet). ECHO was designed to centralize both types of data in a unified computerized process.

After each session of care, the procedures performed are entered by the student in ECHO and validated by the practitioner/teacher who supervised the session of care. The practitioner/teacher assigns a qualitative grade and an autonomy grade for each procedure (from A to C); he/she can also provide a general appreciation of the care session in the free text field. Each procedure is assigned a different weight in the overall grade. Students' skills acquisition is assessed based on the repeated performance of procedures. The possibility of creating non-billable procedures (including educational procedures), which are not listed in the national procedure codification system, allows for the evaluation of skills that are not necessarily linked to technical procedures. Figure 1 presents the simplified UML class diagram of ECHO.

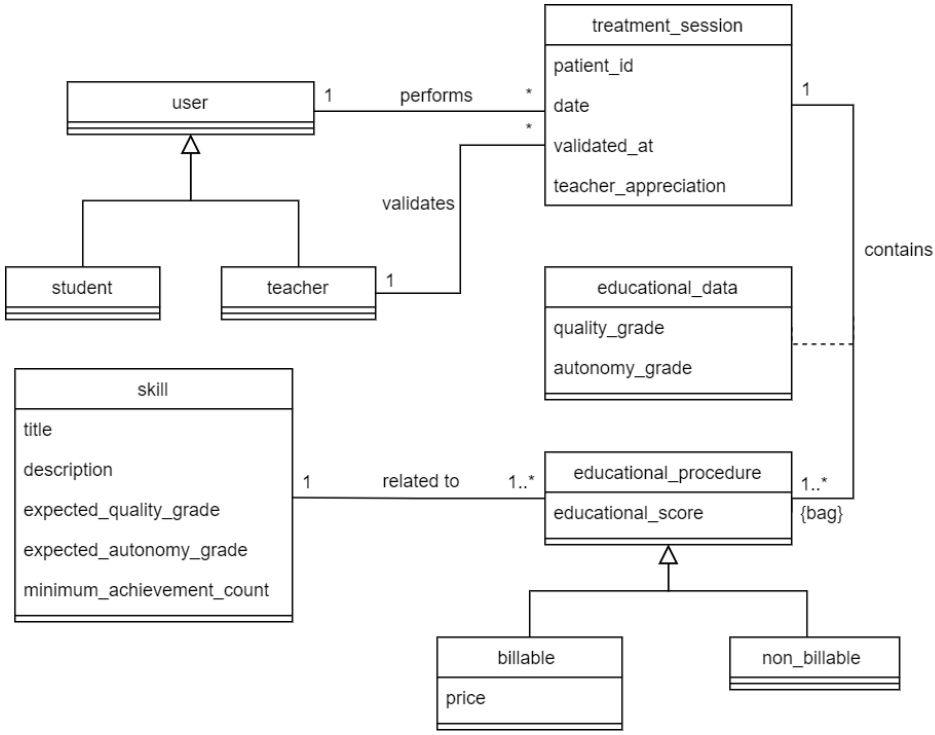


Figure 1. Simplified ULM class diagram of ECHO

ECHO is linked to the hospital information system via dedicated web services, which facilitates the management of user rights access and helps to avoid data entry errors in terms of patient identity. Thanks to this link, users can directly launch and access patients' EDRs from ECHO without extra navigation. Thus, data entry can be performed on ECHO and then in patients' EDRs in a streamlined workflow.

ECHO was launched in 2018/02 and evaluated in late 2018/03 by 102 dental students (40% of all students in the department). Among the respondents, 96% stated that ECHO is easy to use, 86% that it saves time, and 81% that it gives them a better

overview of their activity. Moreover, 79% of respondents declared that ECHO is fairer and more equitable than the internship booklets because it is better at preventing fraud. While many respondents (56%) had mixed or bad opinions about ECHO before it was launched, their perceptions were very positive or positive (98%) by the time of the study. Regarding the difficulties encountered with ECHO, respondents expressed the need for: 1) a tool that facilitates the search for procedure codes in the dropdown list and 2) a tool that reminds faculty members to validate the care session.

4. Discussion

To our knowledge, ECHO is the first information system dedicated to the monitoring and evaluation of dental student activity in a pre-doctoral clinic in France. In our dental department, ECHO has helped to overcome the limitations of the pre-existing health information system which did not account for educational aspects. It has also solved many of the difficulties related to the use of internship paper booklets for student evaluation. Students can now monitor their progress towards their objectives in real time, as can their teachers (figure 2). Those in difficulty can be identified early on and can be offered more sustained support by faculty members. ECHO also makes it possible to track students' skill acquisition throughout their preclinical training.

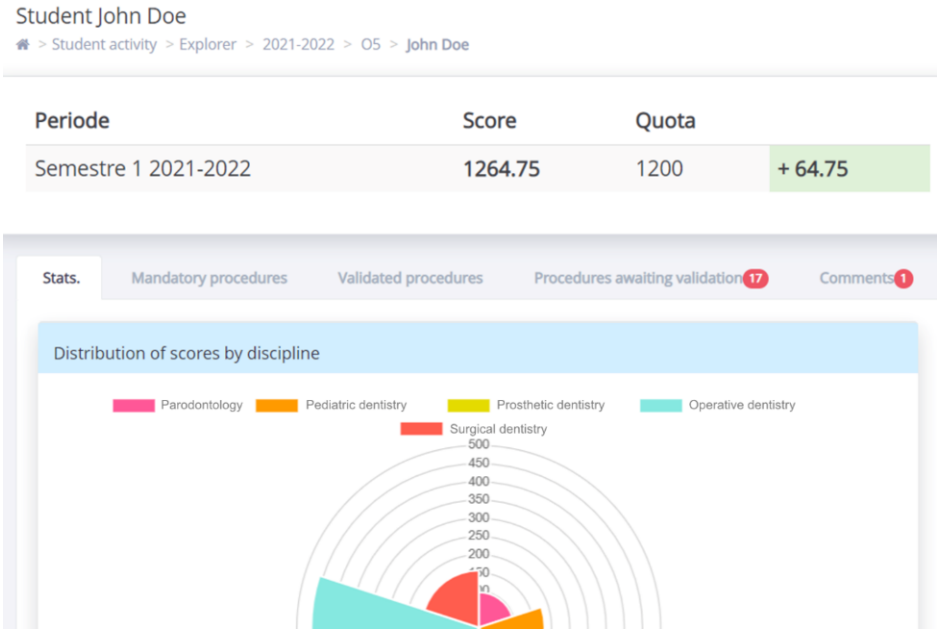


Figure 2 – Example, translated into English, of an available view of student's activity

When building ECHO, we paid special attention to the needs to improve user interface and to facilitate a streamlined workflow, which were identified as a major concern in the literature [9–12]. As regards user interface, we opted for the popular front-end framework Bootstrap because it enables an intuitive usage that reduces student training time. Since Bootstrap is a responsive framework, students started on their own

initiative to use ECHO directly from their smartphone when connected to the hospital's local Wi-Fi network. As regards workflow, we applied the "collect once, use many" principle [13], which allowed to unify administrative and educational data entry in one single process.

The local development and management of ECHO also has advantages in terms of availability and reuse of data. Local practitioners/teachers can reuse the student activity data stored in ECHO for management and research purposes. The inclusion of non-billable educational procedures makes it possible to enter information that is needed for student evaluation, while indirectly enabling better descriptions of patients. In this regard, ECHO may be more useful for clinical research than a registry based on national procedure codes alone.

5. Conclusion

After several years of use, ECHO has fulfilled its educational mission, while also providing a documentary database of the activities of our dental department. The student activity data stored in ECHO are not only directly accessible by faculty members, but can be reused to facilitate departmental management and research and to improve patient follow-up. Future studies are needed to validate the suitability of using non-billable educational procedures to improve the quality of reused data.

References

[1] Institute of Medicine (US) Committee on the Future of Dental Education, Field M. Oral Health Objectives and Dental Education. In: Dental Education at the Crossroads: Challenges and Change. Washington (DC): National Academies Press (US); 1995.
[2] Nalliah RP. Could dental school teaching clinics provide better care than regular private practices? J Investig Clin Dent. 2019;10(2):e12329.
[3] Jeong YN, Natto ZS, Marks ME, Karimbux N. Development and Implementation of a Clinical Dental Faculty Evaluation Instrument. J Dent Educ. 2020;84(3):323–8.
[4] Nalliah RP, Miller K, Stefanac S. Improving Dental School Clinic Operations Using Business Intelligence Data. J Dent Educ. 2019;83(6):614–23.
[5] Kilbourne AM, McGinnis GF, Belnap BH, Klinkman M, Thomas M. The role of clinical information technology in depression care management. Adm Policy Ment Health. 2006;33(1):54–64.
[6] Taylor O. Laravel - The PHP Framework For Web Artisans [Internet]. [cited 2021 Nov 27]. Available from: https://laravel.com/
[7] Mark O, Jacob T. Bootstrap The most popular HTML, CSS, and JS library in the world. [Internet]. [cited 2021 Nov 27]. Available from: https://getbootstrap.com/
[8] Laravel Schema Designer [Internet]. [cited 2021 Nov 27]. Available from: https://laravelsd.com/
[9] Tokede O, Ramoni RB, Patton M, Da Silva JD, Kalenderian E. Clinical documentation of dental care in an era of electronic health record use. J Evid Based Dent Pract. 2016;16(3):154–60.
[10] Sidek YH, Martins JT. Perceived critical success factors of electronic health record system implementation in a dental clinic context: An organisational management perspective. Int J Med Inform. 2017;107:88–100.
[11] Walji MF, Kalenderian E, Tran D, Kookal KK, Nguyen V, Tokede O et al. Detection and characterization of usability problems in structured data entry interfaces in dentistry. Int J Med Inform. 2013;82(2):128–38.
[12] Middleton B, Bloomrosen M, Dente MA, Hashmat B, Koppel R, Overhage JM et al. Enhancing patient safety and quality of care by improving the usability of electronic health record systems: recommendations from AMIA. J Am Med Inform Assoc. 2013;20(e1):e2-8.
[13] Cimino JJ. Collect once, use many. Enabling the reuse of clinical data through controlled terminologies. J AHIMA. 2007;78(2):24–9; quiz 31–2.

Challenges of Trustable AI and Added-Value on Health
B. Séroussi et al. (Eds.)

doi:10.3233/SHTI220582

Using Simulations to Train Medical Students for Unanticipated Technology Failures in Telemedicine

Helen MONKMAN[a,1], Ryan PALMER[b], Shannon IJAMS[c], Liz KOLLAJA[c], Kristin A. RODRIGUEZ[c], Andrew LIEW[c], Anna WICKHAM[b], Frances WEN[c], Bernadette MILLER[c], and Blake J. LESSELROTH[a,c]
[a]*School of Health Information Science, University of Victoria, Canada*
[b]*Kennedy & Company, Washington DC, USA*
[c]*University of Oklahoma-Tulsa, School of Community Medicine, USA*

Abstract. Simulations offer a safe environment for health professional training and the opportunity to predictably and consistently introduce events or variables that may be rare or dangerous in a live setting. Exposing trainees to unanticipated events during simulations can improve their ability to adapt and improvise. The COVID-19 pandemic accelerated the adoption of telehealth worldwide and highlighted the need for better training in health professional schools. In the United States, the Association of American Medical Colleges (AAMC) published new telehealth competency standards in 2021. The AAMC stated that health care providers should be aware of the risks of technology failures, capable of troubleshooting them, and lead systems interventions to improve safety. However, the AAMC does not provide guidance on the specific failures or solutions. In this study, we developed a set of technology failures that can be simulated in a telehealth curriculum. We incorporated one technology failure into a simulated telehealth encounter and gathered students' (N = 53) feedback on the exercise. Students' feedback was overwhelmingly positive. They agreed that integrating technology failures into telehealth simulations provides important practice managing these events during clinical encounters. While telehealth is an important healthcare delivery modality that can improve access-to-care, it is imperative to train medical students to navigate technology failures so that can adeptly manage these issues in clinical practice.

Keywords. Telehealth, telemedicine, simulation, medical education

1. Introduction

Aviation specialists were among the first to recognize the importance of applying human factors and ergonomics principles to create safe, high-reliability systems. Human factors is "the scientific discipline concerned with the understanding of interactions among humans and other elements of a system, and the profession that applies theory, principles, data, and methods to design in order to optimize human well-being and overall system performance" [1]. Given its potential to optimize human and system performance,

[1] Helen Monkman, School of Health Information Science, University of Victoria, Victoria, British Columbia, Canada; E-mail: monkman@uvic.ca

experts across industries – including healthcare – have started adopting human factors principles and methods in product design and professional training.

Clinician training faces similar challenges to airline pilot training. For example, flight assessment programs may not test adaptive skills to manage emergent and unexpected circumstances [2]. Pilots cannot be trained for every possible situation, but they can learn translational skills to manage common situational factors such as startle or surprise [2]. The same applies to medical students and other health professional learners including physician assistants, nursing students, and pharmacy students.

Telehealth has been lauded for its potential to provide health care services to traditionally underserved areas [3]. Telehealth is the use of telecommunications to connect health care stakeholders over a distance [4]. The COVID-19 pandemic accelerated the adoption of telehealth, including virtual home visits, to provide care continuity despite social distancing precautions [5]. However, technology problems are quite common during virtual visits. For example, in one study, participants estimated that nearly one in four (24.4%) encounters needed to be converted to a telephone encounter due to insurmountable technology issues [6].

Recognizing the need for telemedicine practice standards in medical education, the Association of American Medical Colleges (AAMC) published in 2021 a set of 20 telehealth competencies organized into six unique domains [7]. The fifth domain – "Technology for Telehealth" – states that physicians should be able to explain the risk of technology failures, manage (i.e., troubleshoot) them at the point of care, and teach others how to troubleshoot technology failures [7]. However, AAMC does not provide specific examples of technology failures or how providers should respond.

Simulation, or the imitation of real systems, events, and/or processes, is a common human factors method used for training and research. Simulation provides a safe and controlled environment for learners to practice diagnosing and managing unanticipated equipment failures, environmental hazards, and other emergent system properties. For example, one study tested pilots' performance using heads-up displays under low visibility conditions while contending with an obstacle on the runway [2]. In similar fashion, it may be possible to simulate the technology challenges that clinicians contend with including software errors, unstable internet, or patient safety threats (e.g., joining an appointment while operating a vehicle). Because unanticipated events associated with high-stakes tasks can be safely replicated using simulations, we saw an opportunity to apply this concept to telemedicine training.

The purpose of this paper is twofold. First, we share a list of possible technology failures commonly encountered in telehealth and that may be simulated for training exercises. Second, we illustrate how we incorporated a standardized problem into telehealth simulations to test students' abilities troubleshooting technology failures.

2. Methods

2.1. Testable Technology Failures in Telehealth Simulations

To develop a list of technology failures for testing in telehealth simulations, we reviewed the literature on telehealth educational programs, simulations, and implementation reports. Despite the frequency of technology issues in telehealth, surprisingly little has been written about specific failure modes or corrective actions. We did find two studies that describe categories of telehealth technology failures [5,6]. Although these categories

were helpful, they did not provide enough detail to permit modeling during simulation. Therefore, we worked with clinical subject matter experts at the University of Oklahoma-Tulsa School of Community Medicine (OU-TU SCM) to describe technology failures frequently encountered in practice (**Table 1**). Some were suitable for simulation, whereas others would be difficult to implement. For example, we can simulate frozen displays and sound distortion, but it is challenging to standardize and operationalize these kinds of technology issues in a simulation curriculum.

Table 1. Categories of technology failures in the literature and clinician-generated examples of specific and testable strategies suitable for simulation.

Technology Failure	Strategies for Testing Technology Failure
1. Poor video quality	Equip standardized patient's (SP) computer with a low-resolution web camera.
2. Poor audio quality	Modify SP's microphone.
3. Poor internet connectivity (for provider or patient)	Install software to simulate webcam lagging and/or microphone distortion.
4. Software updates delay or prevent use of technology	Instruct SP to stall session while contending with a sham software update. Trigger an on-screen alert notification remotely.
5. Unfamiliarity with technology (provider or patient)	Standardized patient "accidentally" turns off their web camera or sits outside the camera frame.
6. Patients' lack of access to technology or device	Standardized patient connects to virtual session from a public space (e.g., library, coffee shop) to access a computer or Wi-Fi.
7. Lack of integration with electronic medical record	Store-and-forward content such as digital images are not available in the medical record but stored elsewhere on the computer.
8. Connection is not safe or appropriate	Standardized patient connects to session while driving or when there is the threat of domestic violence.

2.2. Simulated Encounter with a Technology Failure

Physician assistant students and medical students (n = 53) at OU-TU SCM participated in a novel telehealth education program as part of a health systems science curriculum. We hosted a one-day workshop on practice management that cross-walked a medical topic (i.e., arterial disease) with a health-systems topic (i.e., telehealth). Faculty provided learners with pre-session readings and delivered a live didactic (a mixture of lectures and cases) on hypertension, dyslipidemia, and peripheral arterial disease. Participants also learned about emerging trends in telemedicine and ways practitioners can transfer clinical communication skills to a telemedicine platform (i.e., "webside manner") [8]. We then introduced the simulated case of a patient presenting to a telemedicine clinic for hypertension management and advised participants to be prepared for technical issues.

We conducted the activity at the school's simulation center. The center has 10 patient examination rooms; each with (1) a stable Wi-Fi connection; (2) landline phones (optional for this exercise, but valuable in the event of actual Wi-Fi issues); (3) videoconferencing software; and (4) a computer workstation with audio and video live-streaming capability. The center employs and trains standardized patients (SP) that can portray a wide range of clinical presentations for the purposes of clinical instruction.

Faculty assigned each student to an exam room with a computer workstation. We furnished students with a written summary of the SPs presenting issue and vital signs. We provided SPs with computers equipped with webcams and microphones; SPs connected from home to the videoconference session. We allocated 25 minutes for each encounter and instructed SPs to turn off their camera mid-way through the session. The

SPs needed to express alarm over the "technical issue". The simulation called for the SP to re-establish a video connection if the student helped them navigate the issue or after several minutes if the student attempted to complete the session without video.

Faculty conducted a 15-minute debriefing session with students after the simulation using a semi-structured script. We asked open-ended questions to identify technical challenges, best-practices, and lessons learned. We also distributed a learner satisfaction questionnaire adapted from an instrument published by Levett-Jones and colleagues [9].

3. Results

Participants' responses on the post-simulation questionnaire indicated that the simulation was successful in several ways (**Figure 1**). For example, most agreed (15; 28%) or strongly agreed (24; 45%) that the simulation was a valuable learning experience. Similarly, most agreed (20; 38%) or strongly agreed (19; 36%) that the simulation helped them manage technical challenges with telemedicine. One participant wrote, "I think this was a good opportunity to practice introductions and navigate technical issues". By contrast, another participant wished for, "more time for patient simulation along with detailed technology troubleshooting solutions". Finally, a third participant wrote, "it was nice to practice being in a situation with technical difficulties present, but I still wouldn't really know what to do if I was in practice and that happened to me."

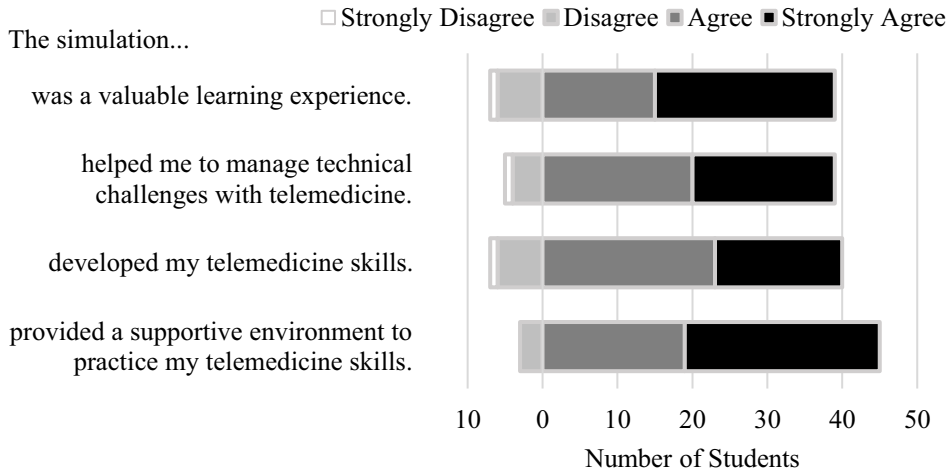


Figure 1. Subset of student responses on the post-session satisfaction survey (N = 53). Neutral responses excluded to best illustrate the valence of responses.

After the workshop, we held a faculty debrief to generate ideas for improving the simulation. Several participants suggested increasing the challenge of the simulation by adding complexity to the clinical case, concealing the possibility of a problem during simulation, or introducing unfamiliar software features. Faculty also suggested that SPs should act more distraught during the camera failure to best portray a patient with low digital literacy and to increase the emotional tension. This could provide students more practice using a range of strategies to calm a patient, de-escalate an encounter, and redirect discussion to the clinical topic. Increasing complexity might create a more naturalistic simulation with better experiential learning opportunities.

4. Discussion and Conclusion

Telemedicine has the potential to improve access to care, empower patients and promote a person-centered experience [10]. However, telemedicine platforms are complex and often challenging for both patients and professionals to use. To provide effective care, providers must be cognizant of the inherent sociotechnical challenges. Educators should use simulations to teach health professional students about commonly encountered failure modes and measure performance when practicing recovery strategies. In this study, we generated a set of technology failures that can be built into existing telehealth training simulations. Results from our satisfaction survey showed that although most participants thought the simulation helped them learn how to troubleshoot basic technology failures, students wanted more time to practice techniques and detailed guidance for addressing specific problems.

The AAMC telehealth competencies emphasize the need for troubleshooting technology failures [7]. Yet, without specific guidance about types of technology failures and corrective actions, educators are left to devise, instrument, and test their own use cases. Having evidence-based and turnkey use cases can help educators create and implement high-quality curricula. Ideally, these technology failures should reflect real-world scenarios and the evolving nature of technology. Thus, more research is warranted to identify types of failures, incidence in practice, consequences, and best-practice recovery methods. Moreover, these findings can be used to inform simulations to ensure health care professionals in training develop adaptive and translational telehealth skills.

References

[1] What is Human Factors and Ergonomics | HFES, (n.d.). https://www.hfes.org/About-HFES/What-is-Human-Factors-and-Ergonomics (accessed January 15, 2022).
[2] Landman A, van Oorschot P, van Paassen MM, Groen EL, Bronkhorst AW, Mulder M, Training Pilots for Unexpected Events: A Simulator Study on the Advantage of Unpredictable and Variable Scenarios, *Hum Factors*. **60** (2018) 793–805. doi:10.1177/0018720818779928.
[3] Gordon HS, Solanki P, Bokhour BG, Gopal RK, "I'm Not Feeling Like I'm Part of the Conversation" Patients' Perspectives on Communicating in Clinical Video Telehealth Visits, *J Gen Intern Med*. **35** (2020) 1751–1758. doi:10.1007/s11606-020-05673-w.
[4] Telehealth | NEJM, (n.d.). https://www.nejm.org/doi/10.1056/NEJMsr1503323 (accessed January 15, 2022).
[5] Mohammed HT, Hyseni L, Bui V, Gerritsen B, Fuller K, Sung J et al. Exploring the use and challenges of implementing virtual visits during COVID-19 in primary care and lessons for sustained use, *PLOS ONE*. **16** (2021) e0253665. doi:10.1371/journal.pone.0253665.
[6] Providers at a Midwestern Academic System Report a Positive Experience with Telehealth During the COVID-19 Pandemic, (n.d.). doi:10.1089/tmj.2021.0410.
[7] Association of American Medical Colleges (AAMC), Telehealth Competencies Across the Learning Continuum, Washington, D.C, 2021. https://store.aamc.org/downloadable/download/sample/sample_id/412/ (accessed January 15, 2022).
[8] Modic MB, Neuendorf K, Windover AK. Enhancing Your Webside Manner: Optimizing Opportunities for Relationship-Centered Care in Virtual Visits, *J Patient Exp*. **7** (2020) 869–877. doi:10.1177/2374373520968975.
[9] Levett-Jones T, McCoy M, Lapkin S, Noble D, Hoffman K, Dempsey J et al. The development and psychometric testing of the Satisfaction with Simulation Experience Scale, *Nurse Education Today*. **31** (2011) 705–710. doi:10.1016/j.nedt.2011.01.004.
[10] Hirko KA, Kerver JM, Ford S, Szafranski C, Beckett J, Kitchen C, et al. Wendling, Telehealth in response to the COVID-19 pandemic: Implications for rural health disparities, *J Am Med Inform Assoc*. **27** (2020) 1816–1818. doi:10.1093/jamia/ocaa156.

Challenges of Trustable AI and Added-Value on Health
B. Séroussi et al. (Eds.)

doi:10.3233/SHTI220583

Software Architecture for Automated Assessment of Prescription Writing

Alireza KHATAMI[a], Anne HOLBROOK[b], Anthony J. LEVINSON[c], Karim KESHAVJEE[a, 1]
[a] *Institute of Health Policy, Management and Evaluation, University of Toronto, Toronto, Canada*
[b] *Division of Clinical Pharmacology & Toxicology, McMaster University, Hamilton, Canada*
[c] *Division of e-learning Innovation, McMaster University, Hamilton, Canada*

Abstract. Prescribing skills are a crucial competency in medical practice considering the increasing numbers of medications available and the increasingly complex patients with multiple diseases faced in clinical practice. Medical students need to become proficient in these skills during training, as required by medical licensing colleges. Not only is teaching the fundamentals of safe and cost-effective prescribing to medical students challenging but evaluating their prescribing skills by faculty members is difficult and time consuming. The COVID-19 pandemic has accelerated the interest in clinically relevant online exams, including automated assessment of short answer style questions. The goal of this project was to design a software to automate the assessment of learners' prescriptions written during low stakes formative assessments. After establishing the components of a legal prescription with multiple medications, and identifying the sources of errors in prescribing and prescribing assessment, we designed and validated an architecture and developed a prototype for automated parsing of learner prescriptions.

Keywords. Online assessment, prescription marking, software architecture, prototype, prescribing skills

1. Introduction

Teaching medical students the art and science of safe and cost-effective prescribing is challenging yet essential for any high quality, sustainable health care system [1]. Aging patients with multimorbidities means that polypharmacy is common and necessary. An increasingly large armamentarium of effective medications means that medical students need to learn how to prescribe many more medications, frequently in combination with other medications. The stakes can be quite high. Increasingly fragile patients end up being hospitalized from iatrogenic causes. More than 700 million prescriptions are written annually in Canada for drugs from approximately 1100 therapeutic groups [2]. Medication errors are common, usually go unrecognized and can pose a serious patient safety hazard [3-7]. Medical students and residents are particularly at risk, with a 7-10% error rate in their prescriptions [4].

[1] Corresponding Author, Dr. Karim Keshavjee, Institute of Health Policy, Management and Evaluation, University of Toronto, Toronto, Canada; E-mail: karim@infoclin.ca

In 2020, the Medical Council of Canada added Prescribing Skills to its specific list of mandatory training objectives for medical schools. However, the number of clinical pharmacology and toxicology specialists who can teach prescribing skills effectively is very small in Canada, making it necessary to maximize use of online resources and make better tools available for teaching prescribing skills.

Multiple choice questions (MCQs) with single best answer are the dominant testing method for most university health professions programs. However, MCQs have well-known limitations, including cueing effect, testing prompted memory rather than ability to generate an answer, and difficulty representing many important areas of medicine [8]. Recently, computer-readable very short answer (VSA) question responses have been validated as a novel pragmatic innovation with excellent results. VSAs better represent actual processes in clinical practice and test actual case-based knowledge and skills with better discrimination than MCQs with single best answer [8,9].

Marking prescriptions and providing high quality feedback to students is time consuming, labor intensive and requires expertise. There is a need for an automated solution that can assist educators to provide high quality feedback to trainees at different stages of training in a timely manner [10]. An automated prescription marker could also create opportunities for more frequent formative assessments and feedback cycles, helping improve prescribing skills faster. This approach has been attempted with some success in the UK and in Australia, but both projects are proprietary, to our knowledge.

Health professions training has relied on in-person, patient-based, preceptor-supervised learning in hospitals and clinics for centuries. However, over the past couple of years, the COVID-19 pandemic has led to the sudden loss of this in-person learning and accelerated the necessity for online curricula, low stakes assessments and high stakes exams. Recognized as one of the most advanced in the world, Canada's medical training system will depend increasingly on high quality eLearning and online evaluation opportunities with appropriate technology and the best faculty trained in delivering education in virtual environments over the next several decades.

2. Methods

Requirements for a Prescription Parser-Marker software were obtained through review of the literature and expert panel consensus. This included parsing of legal prescriptions into their component parts and creating a standardized prescription. A draft software architecture was developed (Figure 1). The Parser-Marker software was designed to handle a variety of question sets with a Marking File that had standard features but was customized for each new question set. This would enable the Parser-Marker to be used for multiple question sets without reprogramming.

A prototype was built using the RapidMiner® (RapidMiner GmbH, Dortmund, Germany) software platform version 9.1 using the built-in Regex functions for text matching. The prototype was tested on responses to 5 clinical scenarios (12 total prescriptions) developed with pre-set marking for alternatives, by an experienced prescribing competence expert (AH). Test scenarios were completed by 10 medical learners using Examplify examination software from ExamSoft (ExamSoft, Dallas, TX, USA) (https://examsoft.com). The prescriptions were typed in manually by students with no prompts or direction about how the prescription should be written. A Marking File was developed using the pre-set marking criteria. Student generated prescriptions were scored by the Parser-Marker software. The prototype-generated scores (N=840 = 10

learners * 12 prescriptions * 7 prescription elements) were examined by two co-authors (KK and AK) to identify issues and reasons why the scores were incorrect.

A root cause analysis was used to identify the reasons for incorrect marking of the prescriptions (Figure 2). Using this information, we redesigned the software architecture to address the issues identified and redeveloped the software to address the issues that lead to incorrect assessments using an automated approach (Figure 3).

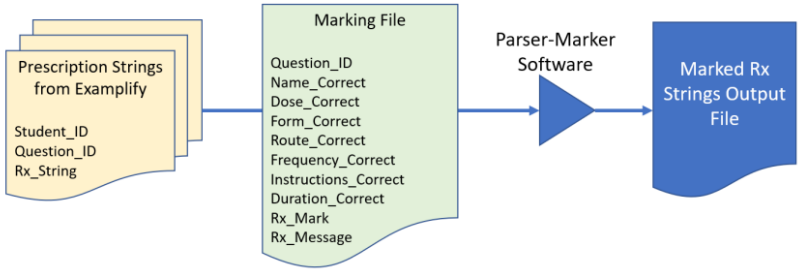


Figure 1. Draft software architecture for prescription marker

3. Root cause analysis of failure of automated prescribing assessment

Several factors affected the performance of the Prescription Parser-Marker. They are grouped into the following factors: student-related, exam software-related, question-related, response-related, marking file-related and software (parsing and matching)-related factors (Figure 2).

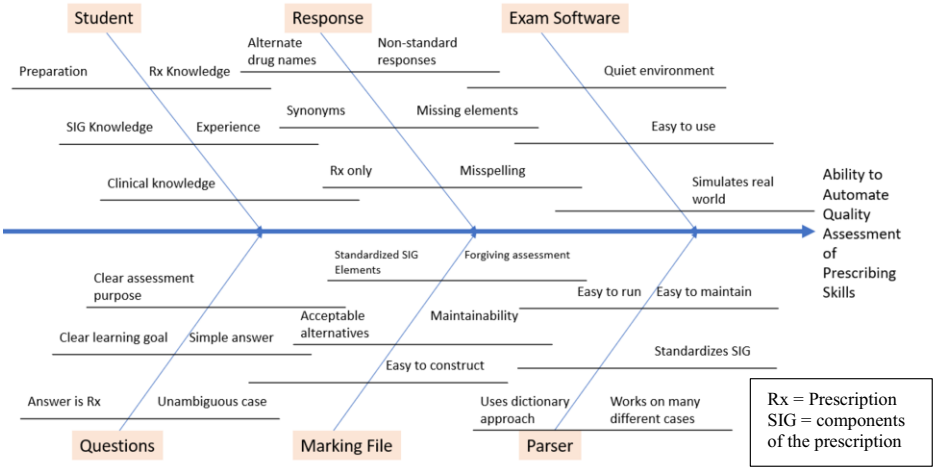


Figure 2. Fishbone diagram of root cause analysis for errors in prescribing assessment

4. Software requirements derived from root cause analysis

The root cause analysis made it clear that variability in the prescription strings (e.g., synonyms for drug names, synonyms for frequency of dosing) was the key issue creating problems with correct marking and accurate scores. Misspellings and non-responses (e.g., to questions which required multiple prescriptions) were other sources of problems.

Variability in students' completion of the prescription elements was a key factor in explaining poor scores on the exam; e.g., it is common for prescribers to leave out certain elements because they are assumed, such as form (e.g., tablet) or route of administration (usually oral route). These are either 'teachable moments' for feedback, or the exam instructions need to be more specific regarding expectations of what constitutes a valid prescription. Variability also arises when there are different possible treatments for a case or different allowed doses or when the correct response is a non-specific instruction (e.g., double the dose of X). Ambiguous questions led to higher variability in answers, threatening successful automation. Straightforward questions help students give unambiguous responses. Ideally, questions need to have a single prescription string as the correct answer. Representing all the potential variability in the Marking File was difficult and it also threatened the ease of creating new Marking Files for new question sets. In addition to being accurate for a single question set, a key requirement for the Parser-Marker software is for it to be easy to extend to new question sets without major reprogramming.

5. Resolution of Architecture for Prototype Prescription Parser-Marker

Variability arising from synonyms can be solved by standardizing the elements of the prescription strings in student responses before being marked. This was accomplished using a 'replace dictionary' (Standardization File) which pre-processes the student response file before passing it onto the Marking File (Figure 3). The Standardization File replaces synonyms with a single, standard word or phrase (e.g., po, by mouth and orally are all replaced with ORAL), making automated marking easier. Since updating the Standardization File is simpler than building variability into the Marking File, it facilitates Marking File generation for new question sets. The Standardization File can

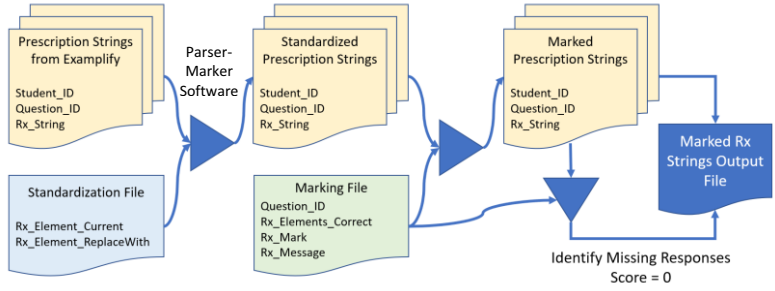


Figure 3. Final architecture for the Parser-Marker Software

be used for all new question sets. Errors in marking due to misspelling of drug names was solved by using the Levenshtein distance of drug names [11]. If the drug name the trainee provided was ≥ 75% distant from the correct response, the system would mark it

0. For example, Atorvastatin and Rosuvastatin have a Levenshtein distance of 75%. Lack of responses to some of the multi-prescription questions was handled by comparing the marked student responses against the original Marking File to identify questions that had not been answered.

Limitations of the current software include the limited pilot testing to date. Future experimental designs are needed for robust evaluation of the developed prototype.

6. Conclusion

Despite increasing use of prescription auto-completer systems, prescription writing competencies are essential parts of medical training and a free-text automated marking system can reduce the current burden on faculty and help students receive timely accurate feedback. We developed a prototype Prescription Parser-Marker which demonstrates a proof of concept for automated prescription writing assessment. The software architecture solves some key issues that threaten automated prescription marking, namely variability in prescription elements, misspellings, and non-responses. The Prescription Parser-Marker is scalable to new question sets by creating a customized Marking File for new question sets. The software can be used outside Canada with minor modification to comply with local requirements.

The next iteration will address: 1) scalability to additional question sets, 2) development of a version control system to better connect question sets to Marking Files, 3) a better user interface to enable users to run the software in their own environments, and 4) improved instructions for additional developers.

References

[1] Holbrook A, Liu JT, Rieder M, Gibson M, Levine M, Foster G, et al. Prescribing competency assessment for Canadian medical students: a pilot evaluation. Can Med Educ J. 2019 Mar 13;10(1):e103-e110. PMID: 30949264; PMCID: PMC6445319.

[2] Canada Health Infoway. Current prescribing and dispensing landscape in Canada: CHI; 2017 [Available from: https://infocentral.infoway-inforoute.ca/en/resources/docs/med-mgmt/1778-report-2017-current-prescribing-and-dispensing-landscape-in-canada. Accessed on 23 Apr 2019].

[3] Baker GR, Norton PG, Flintoft V, Blais R, Brown A, Cox J, et al. The Canadian Adverse Events Study: the incidence of adverse events among hospital patients in Canada. CMAJ : Canadian Medical Association journal = journal de l'Association medicale canadienne. 2004;170(11):1678-86.

[4] Dornan T, Ashcroft D, Heathfield H, Lewis P, Miles J, Taylor D, et al. An in-depth investigation into causes of prescribing errors by foundation trainees in relation to their medical education: EQUIP study. London: General Medical Council. 2009:1-215.

[5] Lazarou J, Pomeranz BH, Corey PN. Incidence of adverse drug reactions in hospitalized patients: a meta-analysis of prospective studies. Jama. 1998;279(15):1200-5.

[6] Ryan C, Davey P, Francis J, Johnston M, Ker J, Lee AJ, et al. The Prevalence of Prescribing Errors amongst Junior Doctors in Scotland. Basic and Clinical Pharmacology and Toxicology. 2011;109:35.

[7] Young H. Lack of pharmacological training causes overuse and misuse of drugs. CMAJ : Canadian Medical Association journal = journal de l'Association medicale canadienne. 2008;178(3):276.

[8] Sam AH, Hameed S, Harris J, Meeran K. Validity of very short answer versus single best answer questions for undergraduate assessment. BMC Med Educ. 2016;16(1):266.

[9] Sam AH, Field SM, Collares CF, van der Vleuten CPM, Wass VJ, Melville C, et al. Very-short-answer questions: reliability, discrimination and acceptability. Med Educ. 2018;52(4):447-55.

[10] Mucklow J, Bollington L, Maxwell S. Assessing prescribing competence. Br J Clin Pharmacol. 2012;74(4):632-9.

[11] Wikipedia contributors. "Levenshtein distance." *Wikipedia, The Free Encyclopedia*. Wikipedia, The Free Encyclopedia, 26 Jan. 2022. Web. 26 Jan. 2022.

Challenges of Trustable AI and Added-Value on Health
B. Séroussi et al. (Eds.)

doi:10.3233/SHTI220584

"ai4health" - Development and Conception of a Learning Programme in Higher and Continuing Education on the Fundamentals, Applications and Perspectives of AI in Healthcare

Daniel KALTHOFF [a], Marieke PRIEN [a] and Nina-Alexandra GÖTZ [1 a]
[a] *GesundheitsCampus Osnabrück, Osnabrück University of AS, Germany*

Abstract. AI applications play an increasingly important role in all areas of healthcare. Therefore, a basic understanding of AI technology for health professionals seems necessary. However, to date there is no learning programme in Germany that includes technological basics, applications, and perspectives of AI in healthcare for interdisciplinary health professions. The *ai4health* project investigates which basic knowledge and competences health professionals need to acquire for an informed handling of AI applications in healthcare, and what the appropriate didactic approach is. Through the qualitative research by interviews and a workshop, six relevant areas of competences were identified. The two most important areas are ELSA and relevant AI applications. Explainability was also highlighted as an important point. The implementation of the topics in a blended learning course for interdisciplinary health professionals and educators in the healthcare sector is now planned.

Keywords. Artificial Intelligence, continuing education, higher education, course development, education and training, future skills, health professionals

1. Introduction

The use of artificial intelligence (AI) entails far-reaching changes in future healthcare. The acquisition of basic AI skills for health professionals therefore seems necessary for the informed handling of AI applications [1,2]. In Germany there exist some AI courses for physician training [3], but to date there are no courses aimed at other health professionals such as nurses, therapists, or administrative staff. A nationally accessible learning programme encompassing the technological basics, applications, and perspectives of AI in healthcare is therefore still missing in Germany.

The project *ai4health* develops a learning programme of this kind. This course is aimed at three target groups: executives in the healthcare sector, health professionals in patient care and administration, as well as teaching staff in the respective fields. The course can be used in line with continuing education for health professionals as well as

[1] Corresponding Author, Nina-Alexandra Götz, GesundheitsCampus Osnabrück, Osnabrück University of Applied Sciences, PO Box 1940, 49009 Osnabrück, Germany; E-Mail: nina.goetz@hs-osnabrueck.de

courses in the corresponding study programmes. The blended learning format includes a MOOC (Massive Open Online Course), SPOC (Small Private Online Course), and attendance phases. The MOOC is hosted on the e-learning platform *AI Campus* (www.ki-campus.org), which offers various learning opportunities for topics around AI. At the same time, teaching materials licensed as OER (open educational resources) are selected or newly created and made available for teaching staff to be used in the SPOC and attendance phase.

To meet the requirements of the interdisciplinary target groups of the course, relevant areas of competence had to be identified. For this purpose, the following research question was investigated: What are the relevant topics and areas of competence health professionals need to acquire to be able to understand the basics, applications and perspectives of AI in healthcare, and how should the knowledge be conveyed?

2. Methods

To investigate the research question, qualitative expert interviews and a workshop with university educators were conducted. Before the interviews and workshop, the *ai4health*-team used literature (e.g. [1,4,5]) as well as media content such as public panel discussions on AI and healthcare (e.g. [6]) to select preliminary topics and subtopics for the course. The selected main topics are: Fundamentals of AI; the role of data; AI applications in healthcare; perspectives of different professional groups; ELSA (ethical, legal, and social aspects); prospects of AI in healthcare. The team created a digital conceptual map containing all topics and subtopics.

Individual interviews were conducted with three students from the following areas: management in healthcare, teaching at vocational schools (health sciences), and nursing management. Subsequently, seven individual interviews and three focus group interviews were conducted with experts from the fields of computer science, health informatics, medicine, law, cognitive science, and speech therapy. In addition, a workshop (n= 10) was held with experts from the fields of nursing and midwifery science. All participants were active in a scientific context or students at one of the two Osnabrück universities.

For the interviews, the experts had received an introduction of the course concept and the digital conceptual map ahead of time, so they were able to get familiar with it and start collecting thoughts on a digital board. The interviews were started off with a reiteration of the concept and preliminary course contents. Then, the experts were asked to discuss the key questions from the point of view of their expertise. During the interview, successive extensions and modifications were made to the conceptual map. The interviews followed a semi-standardized guide that included questions about the thematic content of the programme as well as the didactic implementation. The key questions were further specified depending on the expertise of the interview participants. The following key questions are examples from the guide:

- *Which AI competences do you consider crucial for health professionals?*
- *Do you know of AI applications in your field that are particularly interesting?*
- *How do you rate the relevance of the suggested topics? Do you think there are redundancies or missing topics?*
- *What could be a suitable didactic concept for the implementation of the course and individual course contents?*

In the workshop, the *ai4health*-team introduced the course concept and main topics to the participants. The attendees were split into three groups in which they discussed the following questions amongst each other with guidance from the *ai4health*-team:

- *What chances and risks do you see regarding the use of AI in healthcare?*
- *What are specific topics, questions, or applications that should be included in the course?*
- *What should be the weighting of the topics overall – i.e., should the focus lie on technological basics, data, applications, or ELSA?*

The interviews and workshop took place between mid-July and mid-September 2021. While the interviews were conducted online, the workshop took place on site. All interviews and workshop lasted an average of one hour. Essential verbal content was simultaneously transcribed or documented by transcript.

The evaluation of the interviews is based on a qualitative content analysis [7] using a deductive category system that originates from literature on healthcare and AI. The categorization was carried out by two scientists who made an assignment independently of each other, also to ensure intersubjective traceability. Inconsistencies in categorization were resolved by consensus.

3. Results

The deductive category system in the form of competence areas and topics was validated and partly supplemented inductively. Different competence requirements and topics were highlighted from the experts. The identified areas and subcategories are presented below (see Table 1):

Table 1. Areas of competence and topics for the learning programme "ai4health".

1.	Basics of AI	1.1.	Demarcation and definition of central concepts
		1.2.	Symbolic and subsymbolic methods of AI
		1.3.	Explainable AI
		1.4.	Use cases and underlying AI concepts
	Expected learning outcomes: The participants are able to name basic terms, AI-based procedures and concepts.		
2.	Data management	2.1	Quality of data for big data, AI research and applications
		2.2	Basic concepts of data protection and data security
	Expected learning outcomes: The participants are able to recognize the relevance of data quality for AI processes and know the basic characteristics of data protection and data security in AI applications.		
3.	Applications of AI	3.1.	Health promotion and prevention
		3.2.	Diagnostics and treatment
		3.3.	Research
		3.4.	Care/Nursing
		3.5.	Administration and management of healthcare facilities
		3.6.	Education and training of health professionals
	Expected learning outcomes: The participants are familiar with exemplary AI applications in different areas relevant to health care.		
4.	Profession-specific perspectives	4.1.	Case studies for different professional groups
		4.2.	Role of people / changes in the professional field
	Expected learning outcomes: The participants can explain the potential importance of AI applications for different professions and their possible implications for occupational fields.		
5.	Ethical, legal and social aspects	5.1.	Ethical and social implications
		5.2.	Legal aspects
		5.3.	AI in public discourse

	Expected learning outcomes: The participants are able to deal with possible ethical, legal and social aspects of AI usage in healthcare.		
6.	Future Prospects	6.1.	Current research and technological developments
		6.2.	Incentives, impact and adoption in the healthcare system
		6.3.	Future competence requirements for health professionals
	Expected learning outcomes: The participants can estimate future developments in AI research, technology, and competence requirements. Furthermore, they know basic requirements for the implementation of AI applications in the health sector.		

The experts pointed out that there is no necessity for in-depth knowledge of AI technology or programming, but a necessity for intrapersonal and interpersonal competences. One important intrapersonal competence is the professionals' ability to explain technological knowledge "formula-free" to patients if treatment involves AI applications. Regarding interpersonal competences, the need for interdisciplinary cooperation and a "common language" was highlighted. Didactically, the course should focus on conveying a basic understanding of AI methods by means of exemplary AI applications from the participants' occupational area. Not only does this strengthen the participants' interest in the topic, but it also enables them to assess and evaluate AI applications in their professional life. Furthermore, it dissolves the so-called "black box" of AI especially in medical decision support, and therefore contributes to explainable AI.

In the workshop, "AI applications in healthcare" was deemed the most important topic, followed by ELSA and the AI fundamentals. In concordance with the interviews, attendees of the workshop emphasized that fundamentals (e.g., knowing what makes an algorithm "intelligent"), as well as being able to evaluate AI applications and discussing ethical, legal, and social aspect is more important than knowing technological details. Nevertheless, a basic technological knowledge of AI methods is considered necessary. The workshop participants also highlighted the need for skills to explain AI applications to patients as well as the importance of knowing legal aspects (e.g., who owns the data). In addition to a variety of risks (e.g., data misuse), the participants also saw the benefits of AI applications such as work support or help with personnel deployment planning. Overall, all workshop participants agreed that AI competences are necessary

The results of the qualitative research are now used to develop and test the course and its teaching materials.

4. Discussion

The qualitative research has shown that intrapersonal and interpersonal AI competences should be conveyed in a manner that is user-oriented and closely related to the participants' professional life. This can be achieved by explaining AI methods using exemplary applications from the participants' occupational area. Another focus must be put on ethical, legal, and social aspects.

Accordingly, the weighting of the topics in the course was made in favor of more application references and for ELSA. As for the fundamentals, technological content was reduced while the very basics such as definitions of common terms was increased.

In addition, there is a need for AI competence development tailored to health professionals from the various disciplines. This has already been highlighted in other studies [1]. It has been shown that to promote trust in AI applications, it is necessary to deepen the AI education in both patients [8] and users [9].

However, the difficulty in the conception and development of such a learning programme consists mainly in the heterogeneous target group and limited mathematical and technological skills, especially a lack in programming skills, of the participants.

A limitation of the findings is that other healthcare groups, such as additional therapeutic professionals or managers of health facilities, have not yet been included. Furthermore, only experts from an academic context were interviewed. A progressive validation with employees in healthcare might have produced further results. However, since a conscious use of AI has hardly arrived in everyday working life in the German healthcare organizations, this has so far not been conducted.

From the results of the qualitative part of the study, a learning programme is now being developed that is aimed at interdisciplinary professional groups from the healthcare sector.

5. Conclusions

The development of AI competences should be integrated into the future training and continuing education of health professionals. It is necessary to find the right ratio of technological knowledge and ethical, legal, and social aspects. Furthermore, intrapersonal and interpersonal competences, such as interdisciplinary communication, should be part of the training and continuing education of health professionals.

Funding

This study was funded by the German Federal Ministry of Education and Research (BMBF) (grant: ai4health / 16DHBQP026).

References

[1] EIT Health, McKinsey & Company. Transforming healthcare with AI. The impact on the workforce and organisations. 2020. [cited 2022 Jan 04]. Available from: https://www.mckinsey.com/~/media/McKinsey/Industries/Healthcare%20Systems%20and%20Services/Our%20Insights/Transforming%20healthcare%20with%20AI/Transforming-healthcare-with-AI.ashx

[2] Matheny M, Israni ST, Ahmed M, et al. Artificial Intelligence in Health Care: The Hope, the Hype, the Promise, the Peril. Washington, DC.: National Academy of Medicine; 2019.

[3] Mosch L, Back D-A, Balzer F, et al. Lernangebote zu Künstlicher Intelligenz in der Medizin. Berlin: KI-Campus; 2021.

[4] Sapci A H, Sapci H A. Artificial Intelligence Education and Tools for Medical and Health Informatics Students: Systematic Review, JMIR Med Educ. 2020. 6(1): e19285. DOI: 10.2196/19285: 10.2196/19285.

[5] Plattform Lernende Systeme (eds.): Lernende Systeme im Gesundheitswesen – Bericht der Arbeitsgruppe Gesundheit, Medizintechnik Pflege, München; 2019.

[6] HIMSS. Panel Discussion - Precision Medicine & AI. 2018 [cited 2022 Jan 11]. Available from: https://www.youtube.com/watch?v=UFHsFNznvKg

[7] Mayring P. Qualitative Inhaltsanalyse. Grundlagen und Techniken. 12. edition. Weinheim: Beltz; 2015.

[8] Yang K, Zeng Z, Peng H, et al. Attitudes of chinese cancer patients toward the clinical use of artificial intelligence, Patient Preference and Adherence. 2019. 13: 1867–1875. DOI: 10.2147/PPA.S225952.

[9] Horgan D, Romao M, Morré SA, et al. Artificial Intelligence: Power for Civilisation – and for Better Healthcare. Public Health Genomics. 2019. 22:145–161. DOI: 10.1159/000504785.

Challenges of Trustable AI and Added-Value on Health
B. Séroussi et al. (Eds.)

doi:10.3233/SHTI220585

SNIK Quiz: A Multiple Choice Game About Information Management in Hospitals

Konrad HÖFFNER[1], Arne ROSZEITIS, Max Niclas WÄCHTLER, Franziska JAHN, Alfred WINTER
IMISE, Leipzig University, Germany

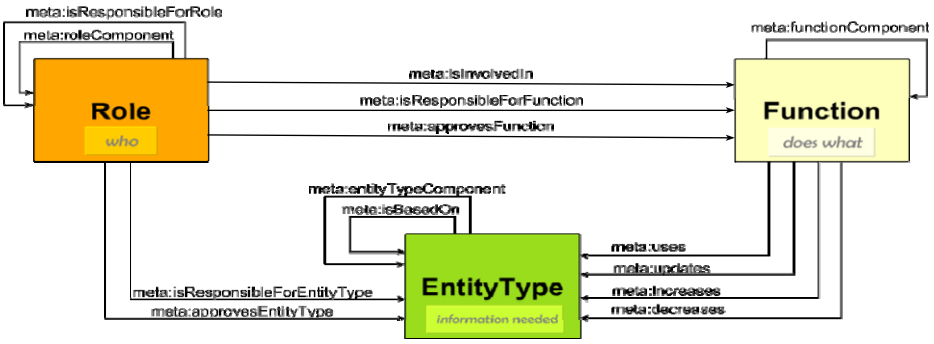
Abstract. SNIK is a knowledge base about the management of health information systems generated by extracting Linked Data from textbooks and other sources. SNIK describes functions, roles executing these functions, and entity types, the information used or updated by these functions. We present SNIK Quiz, a browser game in which students answer multiple-choice questions about information management in hospitals based on SNIK. The questions are semi-automatically generated using templates in order to train basic facts, more complex patterns, and connections between textbooks encoded in SNIK.

Keywords. linked open data, information management, multiple choice quiz

1. Introduction

Medical informatics students, who are trained for executive positions in information management (IM) departments of healthcare institutions, such as hospitals, need a clear terminology of their domain. This terminology is offered by SNIK [1,2], the Semantic Network of Information Management in Hospitals ("Krankenhaus" in German), which integrates knowledge extracted from textbooks [3,4,5] and other sources such as interviews in the form of Linked Open Data. In order to specify, which information should be extracted and to facilitate comparisons, we use a common data model. Because processed textbooks contain abstract knowledge instead of information about any specific hospital, all concepts are modelled as classes and instances simultaneusly, using OWL punning. We thus call our data model the "meta model" in accordance with the term's definition as a shared modelling language [4, p. 8]. 3182 entity types. This paper presents SNIK Quiz, an open source[2] browser game, seeThe central entities of the meta model are enter- prise functions, roles executing these functions, and entity types, the information used or updated by these functions, see Figure 1. SNIK version 1.3.0 contains 81499 triples de- scribing 65 properties and 4719 classes, of which there are 260 roles, 1310 functions and Figure 2, in which students answer multiple-choice questions about health information management (HIM), including an evaluation by domain experts and a student

[1] Corresponding author: Konrad Höffner, IMISE, Leipzig University, Härtelstraße 16–18, 04107 Leipzig, Germany; E-mail: konrad.hoeffner@imise.uni-leipzig.de.



https://www.snik.eu/quiz.

Figure 1. The SNIK meta model

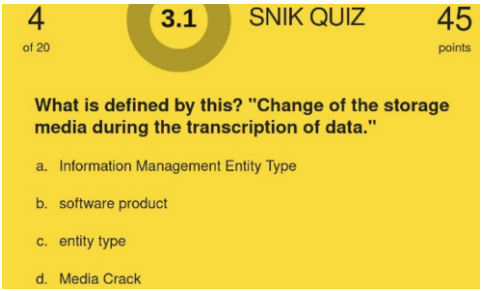


Figure 2. The SNIK Quiz browser game available at

2. Methods

SNIK Quiz is an English single-player game for students of Medical Informatics. This group is diverse and includes students without a background in Semantic Web technologies. Target devices are PCs, tablets and smartphones, so multiple operating systems and input methods need to be supported. SNIK Quiz is designed as an multiple choice quiz with at most four possible answers and published as an open source web application. We employ an approach similar to Clover Quiz [6], which shows that Linked Data knowledge bases can be used to semi-automatically generate cross-domain multiple-choice questions on DBpedia [7]. For each question (*stem*), we generate a correct answer (*key*) and one or more incorrect answers (*distractors*). The data is semi-automatically generated using SPARQL queries on SNIK, such as Listing 1, followed by minimal postprocessing to achieve more natural looking questions. The graph structure of SNIK allows the generation of difficult distractors by using entities semantically close to the key, using path lengths of at most 2 in the graph. Evaluation uses semi-structured interviews with standardized questions but partly open answers.

Listing 1: SPARQL query generating the key and distractors for the *definition* questions.

```
SELECT SAMPLE(REPLACE(STR(?def),STR(?cl),"X","i") AS ?def) SAMPLE(STR(?cl
    ) AS ?cl) SAMPLE(STR(?a1l) AS ?a1l) SAMPLE(STR(?a2l) AS ?a2l) SAMPLE(
   STR(?a3l) AS ?a3l) {
  ?class a owl:Class; rdfs:label ?cl.
  FILTER(LANGMATCHES(LANG(?cl),"en"))
  ?class skos:definition ?def.
  FILTER(STRLEN(?def)>10&&STRLEN(?def)<600).
  FILTER(LANGMATCHES(LANG(?def),"en"))
  ?class (!(meta:subTopClass|rdf:type)){1,2} ?a1,?a2,?a3.
  owl:Class ^a ?a1,?a2,?a3.
  FILTER(?class!=?a1&&?class!=?a2&&?class!=?a3&&?a1<?a2&&?a2<?a3)
  ?a1 rdfs:label ?a1l. FILTER(LANGMATCHES(LANG(?a1l),"en"))
  ?a2 rdfs:label ?a2l. FILTER(LANGMATCHES(LANG(?a2l),"en"))
  ?a3 rdfs:label ?a3l. FILTER(LANGMATCHES(LANG(?a3l),"en"))
} GROUP BY ?class LIMIT 1000
```

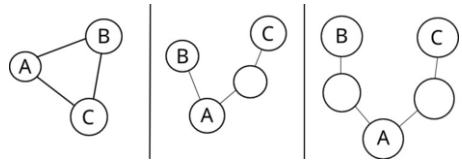


Figure 3. Left to right: correct answer, difficult and easy distractor for the *intertwined* question type [8].

Expections for "good" simple and complex questions are asked. Then, each question category is rated on a scale of 1 to 10. Each question is commented based on the individual critieria of the first step. Finally, the subjective influence is noted on: length of the question and answers, interconnectedness of the content and semantic similarity of the key and distractors.

3. Results

The resulting question data is generated by templates, see Table 1, which are designed to either train basic facts or more complex patterns. The question types definition and subject were presented in a first prototype to students of Medical Informatics during the international Frank van Swieten lectures in 2019 and were positively received. An extended version of SNIK Quiz including the query types definition, definitions, subject, intertwined, closeMatch and occurrence was rated by two experts of hospital information management and one student of Medical Informatics, see Table 2.

4. Discussion

Like Clover Quiz [6], SNIK Quiz contains basic questions that refer to singular facts. These questions were received positively during the Frank van Swieten lectures and the evaluation ratings on subjective complexity show promising results.

Table 1. Quiz question templates with examples. By default distractors are neighbors of degree at most 2.

Template	Description
Subject	Ask for the subject of a (subject, relation, object) triple given the relation and object. Stem *Who is responsible for Medical Admission?*
Key	*Physician*
Distractors	*Surgeon*, *Health Care Professional*, *Senior Physician*
Object	Ask for the object of a (subject, relation, object) triple given the subject and relation. Stem *The Specification Team is responsible for. . .*
Key	*Functional Specification Document*
Distractors:	*Project Team Member*, *Sponsor*, *Defining Project Organization*
Definition	Ask for the entity that fits the given textbook definition.
Stem	*What is defined by this? "X assures a defined quality of all processes and outcomes of the hospital"*
Key	*Internal Quality Management*
Distractors	*Activity*, *Complaint*, *Diagnosis*
Definitions	Present different definitions and ask, which of them fits the given entity. Stem *What defines the term Data Warehouse System?*
Key	*Application component that contains data which have been extracted from other application components, in order to support either hospital management or clinical research.*
Distractors	*Defines the hospital's long-term strategic goals*, *An application component where the controlling rules for data processing are implemented as executable software*, *Summarizes monitored key performance indicators (KPIs) and compares them to the expected future state*
Intertwined	The correct answer and distractors are connected as shown in Figure 3.
Stem	*In the context of Strategic Information Management, which one of the following triples belongs together the most?*
Key	*Chief Information Officer – Department of Information management – IM Staff*
Distractor 1	*Chief Information Officer – Strategic Gap – Corporate Strategy*
Distractor 2	*Chief Information Officer – Ticket Evaluation – Project Monitoring*
Occurence	Ask whether a given term is defined in one of the textbooks [3,4], both or neither [8]. Stem *In which contexts does the term "Health Insurance Company" occur?*
Key	*Strategic Information Management*
Distractors	*Tactical Information Management, Both Contexts, Neither*
CloseMatch	Transfer knowledge about one textbook [3] to the other [4] by confirming or denying statements about entity pairs that are marked as near equivalent [8].
Stem	*In the Strategic Information Management, the Consultant is associated with the Long-Term HIS Planning, while in the Tactical Information Management, the Consultant is responsible for the Functional Specification Document.* (Key: True, Distractor: False)

Questions that test the understanding of relationships involving multiple facts have been successfully eval- uated for middle school questions using an ontology of biology [9], however the style of questions was deemed to be repetitive by two of the evaluating teachers. SNIK Quiz uses multiple templates in order to achieve a more varied question style. A limitation of

Table 2. Interviewee rating on subjective complexity of the questions on a scale between 1 and 10 [8].

	definition	definitions	subject	intertwined	closeMatch	occurence
interview 1	4	3	5–6	7	8	7
interview 2	6	4	7	7–8	7–8	5
interview 3	3	3	6	8	8	7

SNIK Quiz is that qualitative feedback on complex questions is mixed. One of the experts questioned the didactic value of the *intertwined* questions, as the users cannot know the exact graph structure of SNIK and thus may have difficulties deciding, which triplets of concepts are most strongly connected. While the complexity of closeMatch was rated the highest, it was seen as mostly grammatical rather then caused by relationships between entities. Contrary to the initial assumption, shorter questions were preferred as they could be read more quickly. Questions should also clarify that the least specific correct answer is expected in cases such as the example for the subject template in Table 1: As a subclass of the correct answer “physician”, each “senior physician” is also implicitly responsible for medical admission and for “health care professionals” (superclass of physician) this is at least partly true. As SNIK can only cover the domain of HIM as described by the underlying textbooks, distractors may actually be correct answers in the real world.

5. Conclusions

SNIK Quiz shows that HIM knowledge can be used to semi-automatically generate multiple-choice questions. Analysing questions also helps to uncover and fix missing and incorrect facts in SNIK. While preliminary evaluations show promising results, complex question types need to be investigated further. Future work should also include a quantitative evaluation of learning efficiency when supplementing courses with SNIK Quiz.

References

[1] Jahn F, Schaaf M, Paech B, Winter A. Ein Semantisches Netz des Informationsmanagements im Kranken- haus. In: Plödereder E, Grunske L, Schneider E, Ull D, editors. Informatik 2014. Lecture Notes in Infor- matics. Bonn: Gesellschaft für Informatik e.V.; 2014. p. 1491-8.

[2] Höffner K, Jahn F, Kücherer C, Paech B, Brigel S, Schöbel M, et al. Technical Environment for Developing the SNIK Ontology of Information Management in Hospitals. Studies in Health Technology and Informatics. 2017;243:122-6.

[3] Winter A, Haux R, Ammenwerth E, Brigel B, Hellrung N, Jahn F. Health Information Systems: Architectures and Strategies. Health Informatics. Springer London; 2011.

[4] Ammenwerth E, Haux R, Knaup-Gregori P, Winter A. IT-Projektmanagement im Gesundheitswesen. 2nd ed. Stuttgart, Germany: Schattauer; 2014.

[5] Heinrich LJ, Riedl R, Stelzer D. Informationsmanagement: Grundlagen, Aufgaben, Methoden. De Gruyter; 2014.

[6] Vega-Gorgojo G. Clover Quiz: A trivia game powered by DBpedia. Semantic Web. 2019;10(4):779-93.
[7] Lehmann J, Isele R, Jakob M, Jentzsch A, Kontokostas D, Mendes PN, et al. DBpedia—A large-scale, multilingual knowledge base extracted from Wikipedia. Semantic Web. 2015;6(2):167-95.
[8] Roszeitis A. Automatische Generierung komplexer Fragen zum Informationsmanagement auf der Basis der SNIK-Ontologie (German) [Bachelor's Thesis]; 2022.
[9] Stasaski K, Hearst MA. Multiple Choice Question Generation Utilizing An Ontology. In: Proceedings of EMNLP 2017. Copenhagen, Denmark: Association for Computational Linguistics; 2017. p. 303-12.

Challenges of Trustable AI and Added-Value on Health
B. Séroussi et al. (Eds.)

doi:10.3233/SHTI220586

Discovering Rules from a National Exam Repository: A Use Case for Data Analysis from Iranian Medical Schools Entry Exam

Hanieh ZEHTAB HASHEMI[a,b] , Somayeh ABEDIAN[c,d], Parvane PARVASIDEH[a,1], Zahra BAHREVAR[a] and Sina MADANI[e]
[a] *National Center of Medical Education Assessment, Ministry of Health and Medical Education, Tehran, Iran*
[b]*Department of Health Informatics, Virtual University of Medical Sciences, Tehran, Iran*
[c]*Department of Information Technology Management, Islamic Azad University, Qazvin, Iran*
[d]*Ministry of Health and Medical Education, Tehran, Iran*
[e] *Department of Health IT, Vanderbilt University Medical Center, Nashville, TN, USA*

Abstract Many methods have been studied to analyze and interpret patterns and relationships that are embedded in the database to discover new knowledge in educational systems. Association rule mining is a type of data mining that identifies relationships among elements of the dataset. However, because these methods often generate various rules including non-significant ones, it is important to identify the most useful rules. Therefore, evaluating and ranking rules has become a topic of interest in the decision-making process in order to represent the level of usefulness of rules. We incorporated Apriori and Eclat algorithms on an educational dataset of a national medical exam in Iran. The aim of this study is to identify the usefulness of the extracted rules. This method can reliably discover new knowledge by interpreting the prioritized rules. The results show that those who have Scored in the highest category, i.e. [407,493], are accepted and who have scored in the lowest category, i.e. [150,236), are not accepted in the exam regardless of others features. Although, the rules that implication Accept=0 occurs, find out with high confidence, due to a large number of samples in this case. The ranking rules show this method is effective in the identification of insignificant rules that have no effect on decision making.

Keywords. Educational data mining, Association rules, Residency Education, Data Envelopment Analysis

1. Introduction

Data mining techniques have been used in educational systems to analyze student behavior. The amount of data that is stored in these environments is huge and merits investing in the analysis and interpretation of the information(1). Due to this, we were able to extract important knowledge about socioeconomic diversity that may have been hidden otherwise in the dataset (2). Discovering relationships among variables in data may reveal hidden concepts that can be used next for making better decisions(3).

[1] Corresponding Author, Parvane Parvazideh, National Center of Medical Education Assessment, Ministry of Health and Medical Education, Tehran, Iran.Email : parvane.parvasideh@gmail.com

Association Rule Mining(ARM) is a technique for discovering relations in the dataset(4). The items that are associated with a given rule constitute an item set and if this item set occurs frequently in the database, then we can refer to it as a safe rule.
Finding all frequently occurring item sets involves searching all possible item sets that are a power set on the set of all items i.e. 2^(n-1) possible item set. Thus, with increasing the number of items in the dataset, the power set grows exponentially. Therefore, there is a need to determine a threshold value such as the support to calculate the frequency of item sets(5).

ARM discovers the total number of rules that have certain criteria such as confidence and support. It is important to rank rules since the decision-maker must select appropriate rules based on the relevant business application(6). The mined rules are ranked based on support, confidence, and domain-related measures(7).

In this paper, we investigated an educational dataset used in a national exam repository (medical science exams) with two Association Rule algorithms; Apriori and Eclat. We also used the Data Envelopment Analysis to rank and evaluate extracted rules. Our method designates the most important rules to obtain knowledge in order to make better decisions for politicians in the Ministry of Health and Medical Education. Such an approach enables the analyst to have a better perception in order to have a better understanding of the educational domain. In the following, ARM including Apriori and Eclat algorithms are discussed.

2. Methods and Materials

2.1. Association rules

Association Rule Mining discover the relationships between items from the set of transactions. The interestingness of associaton rules are measured by support and confidence. Association rules are regarded as interesting if their support and confidence are greater than the user-specified minimum support and minimum confidence(8).
Several algorithms are available for mining frequently item sets. The Apriori algorithm is an algorithm for mining frequent item sets and uses the minimum support criterion to eliminate infrequent itemsets(9). Since the Apriori algorithm is the first algorithm that was proposed in the domain, it has been improved upon in terms of computational efficiency.

The Eclat algorithm is a more efficient and scalable version of the Apriori algorithm. While the Apriori algorithm works in a horizontal sense imitating the Breadth-First Search of a graph, the Eclat algorithm works in a vertical manner just like the Depth-First Search of a graph. This vertical approach of the Eclat algorithm makes it a faster algorithm than the Apriori algorithm(10).

3. Results

The dataset that we used for our study comes from a national medical exam database that contains students' information. The participants of the exam are students who have passed the primary exam and are moving to the next level. This exam is held annually in Iran with 14,000 students competing for top seats. The students must first sit in this admission exam in order to be eligible for entering university and higher-level education.

Students who were absent from the exam are eliminated from our study. Exam score and several other factors, that determine the students acceptance in the test, were the subject of our study. This exam has two steps: first, students have to answer 200 multiple-choice questions. Each correct and incorrect answer earns 3 positive and 1 negative point, respectively. The volunteers who gain at least 150 out of the total raw score, can select multiple fields of study in their favorite order. Participants who were absent are removed from the next step of the exam. Also, people who have not received an acceptable score (at least 150) cannot move to the next step(11). Both of these two groups were eliminated from our study as well.

Table 1. Selected features

No.	Feature name
2	Quota
3	Sex
6	Military
7	Internship
10	Booklet code
11	Score
12	Pre-internship
13	Average
14	Military staff
15	Military staff scholarship
21	Elite quota
22	Army scholarship

Since the national data repository was composed from several local databases it had to be pre-process first. Missing and duplicate values were removed as well. Data types include both integer and string. The final dataset for our study contained information from 7,723 participants who were accepted in the first phase of the national exam. We have conducted our experiments in three steps. First, As(12) has selected a list of features that are important for this study and has shown in table 1. Then, on such smaller dataset, we used Apriori and Eclat algorithms to generate association rules based on a predefined minimum support and confidence values. Then we used Eclat algorithm since it was more suitable for our sequential data compared to the Apriori algorithm. Some of the relationships, that can be discovered by Eclat approach, may not be found by Apriori approach due to the methodological differences in their algorithms. The results Eclat algorithm is shown in figure 1,2 respectively.

The Apriori algorithm was compiled in Weka environment, whereas for Elcat we used a rules package in R programming language. During data pre-processing, all numeric variables were converted into nominal values, and for some like Score we did more classification (4 categories for Score). The last step was to interpret association rules extracted by Eclat algorithm. The minimum support and confidence values set to 0.1 and 0.9, respectively. We extracted rules with the value of 1 in the class label (Accept=1) and the top 15 values are depicted in figure 1. The first rule shows that when Score belongs to [322,407) range and Army scholarship is 1, then Accept=1 with a confidence score of 0.96% , i.e. students who Score belong to this range and don't have Army scholarship, but have been accepted in this exam. The second rule implies that when Score belongs to [322,407) interval, it is also Accept=1. According to the principle of "rules for proof by cases" logic, we can infer that the first rule can be ignored since the second one is suffice. The person who has Score belonging to [322,407) has been accepted with confidence 0.96%. The next rule implies when Score belongs to this range and Army scholarship=1 and Elite quota=1, i.e. students do not have Army

scholarship and Elite quota, then it concludes Accept=1. Subsequent rule shows that by removing unimportant feature like Army scholarship=1, we can infer who has Score in this range and has not Elite quota, and been accepted. Also, we found who has a Score in this range with Quota =1, and has been accepted.

By reviewing these rules with Accept=0 in Eclat, the following results are found: When the Score belongs to [150,236) and Quota=1 then Accept=0. This means those who are Score at a lower level and do not use any Quota, are not accepted in exam. There is another rule that says Score belongs to [150,236) and Quota=1 and Army scholarship=11 then Accept=0. This means even if the Army scholarship is used, the result does not change.

rules	support	confidence
{Score=[322,407),Army.scholarship=1} => {Accept=1}	0.104752	0.964243
{Score=[322,407)} => {Accept=1}	0.108119	0.964203
{Score=[322,407),Elite.quota=1,Army.scholarship=1} => {Accept=1}	0.103587	0.963855
{Score=[322,407),Elite.quota=1} => {Accept=1}	0.106953	0.963827
{Quota=1,Score=[322,407),Army.scholarship=1} => {Accept=1}	0.103457	0.963812
{Score=[322,407),Military.staff=0,Military.staff.scholarship=1,Army.scholarship=1} => {Accept=1}	0.103328	0.963768
{Score=[322,407),Military.staff=0,Army.scholarship=1} => {Accept=1}	0.103328	0.963768
{Score=[322,407),Military.staff.scholarship=1,Army.scholarship=1} => {Accept=1}	0.103328	0.963768
{Score=[322,407),Military.staff=0,Military.staff.scholarship=1} => {Accept=1}	0.106694	0.963743
{Score=[322,407),Military.staff.scholarship=1} => {Accept=1}	0.106694	0.963743
{Score=[322,407),Military.staff=0} => {Accept=1}	0.106694	0.963743
{Quota=1,Score=[322,407)} => {Accept=1}	0.106694	0.963743
{Quota=1,Score=[322,407),Elite.quota=1,Army.scholarship=1} => {Accept=1}	0.102292	0.963415
{Score=[322,407),Military.staff=0,Military.staff.scholarship=1,Elite.quota=1,Army.scholarship=1} => {Accept=1}	0.102162	0.96337
{Score=[322,407),Military.staff=0,Elite.quota=1,Army.scholarship=1} => {Accept=1}	0.102162	0.96337

Figure 1. Apriori rules extracted

```
Best rules found:

 1. Quota=1 Score='(-inf-235.75]' Military staff=0 Army scholarship=11 3144 ==> Accept=0 3053    conf:(0.97)
 2. Quota=1 Score='(-inf-235.75]' Military staff=0 Military staff scholarship=1 Army scholarship=11 3144 ==> Accept=0 3053    conf:(0.97)
 3. Quota=1 Score='(-inf-235.75]' Military staff=0 Elite quota=11 Army scholarship=11 3144 ==> Accept=0 3053    conf:(0.97)
 4. Quota=1 Score='(-inf-235.75]' Military staff=0 Military staff scholarship=1 Elite quota=11 Army scholarship=11 3144 ==> Accept=0 3053    conf:(0.97)
 5. Quota=1 Fighter quota=1 Score='(-inf-235.75]' Military staff=0 Army scholarship=11 3102 ==> Accept=0 3012    conf:(0.97)
 6. Quota=1 Fighter quota=1 Score='(-inf-235.75]' Military staff=0 Military staff scholarship=1 Army scholarship=11 3102 ==> Accept=0 3012    conf:(0.97)
 7. Quota=1 Fighter quota=1 Score='(-inf-235.75]' Military staff=0 Elite quota=11 Army scholarship=11 3102 ==> Accept=0 3012    conf:(0.97)
 8. Quota=1 Fighter quota=1 Score='(-inf-235.75]' Military staff=0 Military staff scholarship=1 Elite quota=11 Army scholarship=11 3102 ==> Accept=0 3012    conf:(0.97)
 9. Fighter quota=1 Score='(-inf-235.75]' Military staff=0 Army scholarship=11 3123 ==> Accept=0 3025    conf:(0.97)
10. Fighter quota=1 Score='(-inf-235.75]' Military staff=0 Military staff scholarship=1 Army scholarship=11 3123 ==> Accept=0 3025    conf:(0.97)
11. Fighter quota=1 Score='(-inf-235.75]' Military staff=0 Elite quota=11 Army scholarship=11 3123 ==> Accept=0 3025    conf:(0.97)
12. Fighter quota=1 Score='(-inf-235.75]' Military staff=0 Military staff scholarship=1 Elite quota=11 Army scholarship=11 3123 ==> Accept=0 3025    conf:(0.97)
13. Quota=1 Score='(-inf-235.75]' Military staff=0 3260 ==> Accept=0 3155    conf:(0.97)
14. Quota=1 Score='(-inf-235.75]' Military staff=0 Military staff scholarship=1 3260 ==> Accept=0 3155    conf:(0.97)
15. Quota=1 Score='(-inf-235.75]' Military staff=0 Elite quota=11 3259 ==> Accept=0 3154    conf:(0.97)
```

Figure 2. Eclat rules extracted

4. Conclusion

In this study, we focused on an educational dataset containing the results of a national medical school entry exam. The repository included information about students themselves as well as exam results. We used existing algorithms to discover patterns and relationships that may have been hidden from the subject matter experts and exam developers. Therefore, leveraging relationship mining techniques is a vital task for uncovering knowledge from existing data and establishing a learning knowledge management cycle. The results show that those who have Scored in the highest category, i.e. [407,493], are accepted and who have a Score in the lowest category, i.e. [150,236), are not accepted in the exam regardless of others features. Although, the rules that implication Accept=0 occurs, find out with high confidence, due to a large number of

samples in this case. The ranking rules show this method is effective in the identification of insignificant rules that have no effect on decision making.

References

[1] Singh A, Rocke S, Pooransingh A, Ramlal CJJIToE. Improving student engagement in teaching electric machines through blended learning. 2019;62(4):297-304.
[2] Gupta A, Garg D, Kumar PJIToE. Analysis of students' ratings of teaching quality to understand the role of gender and socio-economic diversity in higher education. 2018;61(4):319-27.
[3] Swart AJJIToE. Evaluation of final examination papers in engineering: A case study using Bloom's Taxonomy. 2009;53(2):257-64.
[4] Agrawal R, Imielinski T, Swami AJItok, engineering d. Database mining: A performance perspective. 1993;5(6):914-25.
[5] Zhao Y, Zhang C, Cao L. Post-Mining of Association Rules: Techniques for Effective Knowledge Extraction: Techniques for Effective Knowledge Extraction: IGI Global; 2009.
[6] Buchanan RA, Forster DJ, Douglas S, Nakar S, Boon HJ, Heath T, et al. Philosophy of Education in a New Key: exploring new ways of teaching and doing ethics in education in the 21st century. 2021:1-20.
[7] Obata T, Ishii HJEJoOR. A method for discriminating efficient candidates with ranked voting data. 2003;151(1):233-7.
[8] Agrawal R, Imieliński T, Swami A, editors. Mining association rules between sets of items in large databases. Proceedings of the 1993 ACM SIGMOD international conference on Management of data; 1993.
[9] Agrawal R, Srikant R, editors. Fast algorithms for mining association rules. Proc 20th int conf very large data bases, VLDB; 1994: Citeseer.
[10] Zaki MJJItok, engineering d. Scalable algorithms for association mining. 2000;12(3):372-90.
[11] Heidarzadeh A, Zehtab Hashemi H, Parvasideh P, Bahrevar ZJJoM, Cultivation. A Historical Review on the Achievements and Challenges of the Residency exam in a Forty Years Trend (1357-1397) in Iran. 2019;28(3):11-20.
[12] Hashemi HZ, Parvasideh P, Larijani ZH, Moradi F, editors. Analyze Students Performance of a National Exam Using Feature Selection Methods. 2018 8th International Conference on Computer and Knowledge Engineering (ICCKE); 2018: IEEE.

Challenges of Trustable AI and Added-Value on Health
B. Séroussi et al. (Eds.)

doi:10.3233/SHTI220587

Potential Uses of Assistive Robotic Systems in Acute Inpatient Care

Christoph OHNEBERG[a,1], Nicole STÖBICH[a], Angelika WARMBEIN[b], Ivanka RATHGEBER[b], Uli FISCHER[b] and Inge EBERL[a]
[a] *Catholic University of Eichstätt-Ingolstadt, Germany*
[b] *Hospital of the Ludwig-Maximilians-University (LMU) Munich, Germany*

Abstract. Potential uses of assistive robotic systems in acute inpatient care were defined based on the Framework for Complex Interventions developed by the Medical Research Council (MRC). This process of definition requires the consideration of personal-related and contextual factors.

Keywords. Assistive Robotics, Nursing Robotics, Qualitative Research, REsPonSe

1. Introduction

Participative and practice-related approaches to developing application scenarios are being increasingly used in the development of complex robotic interventions [1]. This also applies to the REsPonSe project, whose aim is to develop potential uses for a digital, robotic assistance system for acute inpatient care, in order to alleviate the workload of nursing staff.

2. Methods

The research design was based on the Framework for Developing and Evaluating Complex Interventions of the Medical Research Council (MRC), which includes four phases. Data collection was undertaken in the first phase Developing/Identifying Interventions [2]. Twelve individual episodic-narrative interviews [3] were carried out with nursing and support staff from acute inpatient care. The data were analysed successively and iteratively according to the following coding techniques of Saldaña [4]: Descriptive-, Process-, Initial-, Magnitude-, Values-, Focused-, and Axial-Coding. A validation of the results was carried out in a project workshop.

3. Results

The potential uses of assistive robotic systems can be divided into five areas: nursing and patient-related tasks, digital communication, organizational tasks, documentation and

[1] Corresponding Author, Christoph Ohneberg, Catholic University of Eichstätt-Ingolstadt, Kapuzinergasse 2, 85072 Eichstätt-Ingolstadt, Germany; E-mail: Christoph.Ohneberg@ku.de.

information, and delivery and pick-up activities. Surgical, urological and oncological departments were discussed as possible deployment locations of use. The use of assistive robotic systems in intensive care units or psychiatric departments was considered inappropriate by those interviewed. Furthermore, personal-related and contextual factors such as illness, biography, age, technical ability and stress were identified by interviewees as influencing factors with regard to the development and application of assistive robotic systems.

4. Discussion

The interviewees could all see potential uses for assistive robotic systems in a clinical setting. However, the interviewees evaluated the potential for increased workload resulting from the use of robotic systems critically, for example, due to technical problems. In addition to the robotic system and its potential uses, environmental and contextual factors associated with the different clinical departments, such as stress and disruptions to workflow, can have varying effects on the levels of acceptance of system users [1].

5. Conclusion

In addition to technical aspects such as usability, aesthetics and feasibility of implementation, research and technology development in this area needs to focus more on the initial systemic situation of the relevant location of use with its specific cultural and patient-related factors.

Acknowledgements

The project REsPonSe (Robotic system to alleviate the work of nursing staff by reducing time spent on service activities) is funded by the Sponsorship Programme “Robotics in Nursing Care” by the Federal Ministry of Education and Research (identification number 16SV8416).

References

[1] Mutlu B, Forlizzi J. Robots in organizations: The role of workflow, social, and environmental factors in human-robot interaction. In: Living with robots, Proceedings of the Third ACM/IEEE Conference on Human-Robot Interaction, March 12 - 15, 2008, Amsterdam, The Netherlands ; HRI 2008. New York, NY: ACM Press; 2008. p. 287–94.

[2] Skivington K, Matthews L, Simpson SA, Craig P, Baird J, Blazeby JM, Boyd KA, Craig N, French DP, McIntosh E, Petticrew M, Rycroft-Malone J, White M, Moore L. A new framework for developing and evaluating complex interventions: update of Medical Research Council guidance. BMJ. 2021; 374n2061. doi:10.1136/bmj.n2061.

[3] Flick U. Qualitative Sozialforschung: Eine Einführung. 1st ed. Reinbek bei Hamburg: Rowohlt-Taschenbuch-Verl.; 2017. 617 p.

[4] Saldaña J. The coding manual for qualitative researchers. Los Angeles: Sage; 2021. 414 p.

Challenges of Trustable AI and Added-Value on Health
B. Séroussi et al. (Eds.)

doi:10.3233/SHTI220588

AI Diagnostic Technologies and the Gap in Colorectal Cancer Screening Participation

Saleem AMEEN[a,1], Ming Chao WONG[b], Kwang Chien YEE[a], Christian NØHR[c] and Paul TURNER[b]
[a]*School of Medicine, University of Tasmania*
[b]*School of Information and Communication Technology, University of Tasmania*
[c]*Department of Planning, Aalborg University, Denmark*

Abstract. AI augmented clinical diagnostic tools are the latest research focus in colorectal cancer (CRC) detection. While the opportunity presented by AI-enhanced CRC diagnosis is sound, this paper highlights how its effectiveness with respect to reducing CRC-related mortality and enhancing patient outcomes may be limited by the fact that patient participation remains extremely low globally. This paper builds a foundation to consider how human factors tend to contribute to low participation rates and suggests that a more nuanced socio-technical approach to the development, implementation and evaluation of AI systems that is sensitive to the psycho-social and cultural dimension of CRC may lead to tools that increase screening uptake.

Keywords. colorectal cancer screening, socio-technical design, patient outcomes

1. Introduction

Colorectal cancer (CRC) is the second leading cause of cancer related death in the world [1]. To reduce CRC-related mortality, high-risk citizens are invited to undertake a tiered two-stage screening process of (1) Immunochemical Faecal Occult Blood Test (FOBT) screening with (2) follow-up colonoscopy, that aims to detect early traces of the disease. While this Gold Standard approach has been shown to reduce CRC-related mortality, its effectiveness is dependent on reaching a screening coverage greater than 65-80%, and several high-income nations have failed to reach these targets [2]. For example, in Australia, participation rates have plateaued at ~40% over the last 5 years, and participation in follow-up colonoscopy by positive FOBT patients is also low (50 – 70%) [3]. Concerningly, marginalised groups at highest risk of CRC participate the least in screening. This is despite efforts to raise awareness through (a) mass media public health campaigns, (b) targeted support programs, and (c) primary care engagement and health systems improvement [4]. Several qualitative studies have suggested that CRC screening adoption and adherence is more often driven by complex psycho-social and cultural interactions. Most significantly, fear, anxiety, stigma, shame, or uneasiness associated with a positive cancer diagnosis, or the invasiveness of a colonoscopy, are reported as major barriers preventing screening participation [4,5]. When multi-factorial barriers present, such as time scarcity or inaccessibility to healthcare centres, the participation problem is exacerbated.

[1] Corresponding Author: Saleem Ameen, E-mail: saleem.ameen@utas.edu.au

2. AI and the Need for a Socio-Technical Approach in the Design of New Tools

There has been considerable research on AI systems in the screening and diagnosis of CRC [6]. AI polyp detection systems have emerged, for instance, to respond to the fact that 25% of polyps are missed during routine colonoscopies, which increases the risk of interval cancer and can negate the benefits of screening [7]. However, for such systems to be advantageous, patients must participate in screening. There is a gap in existing knowledge on how AI tools can be developed in ways that are sensitive to the psycho-social and cultural dimensions that prevent uptake in existing screening programmes to begin with. Utilising a socio-technical approach to the development of future AI technologies may lead to a more significant impact on patient outcomes. For example, a socio-technical approach would identify that patient motivations for non-participation perpetuate around themes of cancer anxiety, colonoscopy invasiveness, and accessibility of healthcare interventions. These issues could be mitigated with the right technology, such as utilising AI-augmented capsule endoscopy devices to target the detection of *precancerous* lesions through an at-home consumer-based health delivery model. Notably, there are many candidates for an intervention (blood, urine, stools, mobile images, smart toilets, ubiquitous health data, among others) and identifying the interventional context for AI development that maximises participation is important.

However, AI model efficacy is sensitive to the type of data used and there is an inevitable interplay between user interaction and data. A model that is accurate, but that requires data the average citizen cannot or is unwilling to provide, may lack utility. Furthermore, implicit contextual, cultural, and temporal biases permeate most ML data distributions [8]. Evaluation methodologies that ensure representation of marginalised patients that are most at risk of CRC, are lacking. Our research aims to develop a robust and nuanced socio-technical framework to the design, development, implementation, and evaluation of AI systems in clinical practice.

References

[1] Sung H, Ferlay J, Siegel RL, Laversanne M, Soerjomataram I, Jemal A, et al. Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. CA: A Cancer Journal for Clinicians. 2021;71

[2] Navarro M, Nicolas A, Ferrandez A, Lanas A. Colorectal cancer population screening programs worldwide in 2016: An update. World journal of gastroenterology. 2017 May 28;23(20):3632.

[3] National Bowel Cancer Screening Program monitoring report 2021, Summary [Internet]. Australian Institute of Health and Welfare. Available from: https://www.aihw.gov.au/reports/cancer-screening/nbcsp-monitoring-report-2021/summary

[4] Lotfi-Jam K, O'Reilly C, Feng C, Wakefield M, Durkin S, Broun K. Increasing bowel cancer screening participation: integrating population-wide, primary care and more targeted approaches. Public health research & practice. 2019 Jul 31;29(2).

[5] Llovet D, Serenity M, Conn LG, Bravo CA, McCurdy BR, Dubé C, Baxter NN, Paszat L, Rabeneck L, Peters A, Tinmouth J. Reasons for lack of follow-up colonoscopy among persons with a positive fecal occult blood test result: A qualitative study. The American journal of gastroenterology. 2018 Dec;113(12):1872.

[6] Mitsala A, Tsalikidis C, Pitiakoudis M, Simopoulos C, Tsaroucha AK. Artificial Intelligence in Colorectal Cancer Screening, Diagnosis and Treatment. A New Era. Current Oncology. 2021 Jun;28(3):1581-607.

[7] Kim NH, Jung YS, Jeong WS, Yang HJ, Park SK, Choi K, et al. Miss rate of colorectal neoplastic polyps and risk factors for missed polyps in consecutive colonoscopies. Intestinal research. 2017 07;15(3):4118.

[8] Sambasivan N, Kapania S, Highfill H, Akrong D, Paritosh PK, Aroyo LM. "Everyone wants to do the model work, not the data work": Data Cascades in High-Stakes AI; 2021.

Challenges of Trustable AI and Added-Value on Health
B. Séroussi et al. (Eds.)

doi:10.3233/SHTI220589

Artificial Intelligence Competencies in Postgraduate Medical Training in Germany

Louis AGHA-MIR-SALIM[a], Lina MOSCH[a,b], Sophie A. I. KLOPFENSTEIN[a], Maximilian M. WUNDERLICH[a], Nicolas FREY[a], Akira-Sebastian PONCETTE[a,b,c] and Felix BALZER[a,1]
[a] *Institute of Medical Informatics, Charité – Universitätsmedizin Berlin, Germany*
[b] *Department of Anesthesiology and Intensive Care Medicine, Charité – Universitätsmedizin Berlin, Germany*
[c] *Hacking Health Berlin, Germany*

Abstract. Routine medical care is to be transformed by the introduction of artificial intelligence (AI), requiring medical professionals to acquire a novel set of skills. We assessed the density of AI learning objectives and the availability of courses containing AI content in postgraduate medical education in Germany. The results reveal general paucity in AI learning objectives and content across (sub-)specialty training and continuing medical education (CME) in Germany. Innovative and regulatory solutions are needed to herald an era of physicians competent in navigating medical AI applications.

Keywords. artificial intelligence, postgraduate medical education, CME

1. Introduction

Everyday medical care is on the verge of revolution with the use of big data and introduction of artificial intelligence (AI). Physicians need to become competent in navigating these novel methods as part of modern care [1]. However, the status quo exhibits discrepancies between imparting competencies physicians require and the education they receive on this matter [2]. Our aim was to investigate the prevalence of AI learning objectives as part of medical specialty and subspecialty training and continuing medical education (CME) in Germany.

2. Methods

After ethical approval [3], we assessed training regulations by the German Medical Association (GMA) and all 17 state-level medical associations regarding AI learning objectives within medical (sub-)specialty trainings between April and June 2021. To assess all CME offerings encompassing AI related content throughout Germany, we surveyed state-level medical associations and conducted a search of the GMA register.

[1] Corresponding Author, Felix Balzer, Institute of Medical Informatics, Charité – Universitätsmedizin Berlin, Charitéplatz 1, 10117 Berlin, Germany; E-mail: felix.balzer@charite.de.

3. Results

None of the 56 specialty trainings in Germany exhibit AI-related content. One (1/56, 1.8%) subspecialty training contained AI-related content, namely the subspecialty *Medical Informatics*, where AI is not a standalone subject area. Instead, it focuses on technical foundations of informatics (i.e., programming languages, databases) and application-specific concepts (i.e., decision support systems, reasoning methods). Concerning CME, seven of 17 state-level medical associations currently offer AI-related courses. The remaining 10 do not offer any AI-related courses, seven of which do not plan to do so in the future. 30 of 87,136 (0.03%) CME courses listed are AI-related. Twenty-three of these 30 were specific in specialty and application with general scarcity in teaching basic principles of AI.

4. Discussion

There is a clear mismatch between the prospect of physicians educated in encountering medical AI competently and the current landscape of postgraduate medical curricula in Germany. In lieu, medical training is very much antiquated, neglecting competencies enabling doctors to apply AI in an informed and critical fashion. There needs to be a shift towards anchoring digital skills, e.g., deploying AI and machine learning at the bedside, systematically throughout postgraduate training. Bringing this change about is difficult for several reasons. One challenge is the fragmented decision-making and heterogeneity in the design of (sub-)specialty training curricula given the multitude of state-level medical associations in Germany. These are responsible for implementing AI-related competencies in their training programs – though this implementation process takes time and medical AI research and practice is advancing quickly. In order to upskill practicing doctors, CME trainings on medical AI are in high demand and display an important vehicle in networking stakeholders and offer an ability to scale existing (online) offerings to large audiences. Another challenge lies in teaching relevant AI competencies to physicians with the backdrop of heterogeneous applications of AI throughout specialties.

5. Conclusions

The current landscape of secondary medical education serves as symbol for the general shortcoming in equipping health professionals for future-proof medicine in Germany. Redesigning postgraduate medical education to addresses challenges of the future requires innovative models and top-down regulations to impart AI learning objectives throughout postgraduate medical training underlying the German federal system.

References

[1] Topol E. Preparing the healthcare workforce to deliver the digital future [Internet]. 2019. Available from: https://topol.hee.nhs.uk/wp-content/uploads/HEE-Topol-Review-2019.pdf
[2] Mosch, L., Back, D.-A., Balzer, et al. (2021). Lernangebote zu Künstlicher Intelligenz in der Medizin. Berlin: KI-Campus. DOI:10.5281/zenodo.5497668
[3] Ethikausschuss am Campus Charité - Mitte, Vote: EA1/236/19, Chairperson: Prof. Dr. R. Morgenstern

Challenges of Trustable AI and Added-Value on Health
B. Séroussi et al. (Eds.)

doi:10.3233/SHTI220590

Online Usability Tool - First Experiences from Testing a Ventilator Interface

Beatrice COLDEWEY[a,1], Rainer RÖHRIG[a] and Myriam LIPPRANDT[a]
[a] *Institute of Medical Informatics, Medical Faculty, RWTH Aachen University, Germany*

Abstract. Usability tests of medical devices are mainly conducted on-site, but remote tests can also be suitable for quick feedback. Using the online survey tool SoSci Survey and videos of a ventilator interface prototype, a usability test environment was developed which allows participation independent of time and place.

Keywords. Usability Testing; Medical Device; Online Test Environment

1. Introduction

Human-machine interaction is a safety-critical aspect in the use of medical devices such as ventilators. Continuous data and different types of alarms are to be presented in a comprehensible way. Since medical decisions are made on the content presented, an early analysis of the graphic concepts is essential and required by regulation. Usability tests with users are a suitable method for testing medical devices and providing information on safety and hazards during use (1,2). Due to the Covid-19 pandemic, but also for fast, iterative feedback, remote synchronous or asynchronous tests have become more important to evaluate medical devices (3).

In order to be able to guarantee adequate usability, suitable tools must be available for implementing remote tests. Since the devices or prototypes to be tested are not provided on-site, they must be made available to the user via another modality (e.g. simulation or video). The aim of this work is the development of such an online usability test environment for the implementation of asynchronous remote tests using the example of a ventilator interface.

2. Methods

An enhanced UI-design of highlighting “air trapping”, a pathological mechanism of trapping too much air in the lung, in a ventilator interface was developed. This prototype requires a continuous data stream of ventilatory data depending on patient status and ventilator setting for a realistic presentation. This representation does not exist separately from the ventilator device. Therefore, we have included a simulated prototype as a video in a usability testing environment.

[1] Corresponding Author: Beatrice Coldewey, bcoldewey@ukaachen.de

As usability test environment, the open source web application SoSci Survey (https://www.soscisurvey.de/) was used. The functional scope of SoSci Survey includes qualitative and quantitative question types to measure usability and identify potential comprehension problems in human-machine interaction. It is also a comprehensive tool for planning, conducting and evaluating online surveys with a high level of data privacy.

3. Results

The comprehensibility of the representation of different alarm messages as well as highlighting is to be evaluated. Sosci Survey was configured and extended to implement a cross-over study design. Through the randomised presentation of different interfaces, dedicated usability tasks could be tested. An evaluation of the efficiency was not clearly measurable due to interfering factors, e.g. brief interruption of the task. To evaluate the effectiveness, the test persons were asked to write down the recognised ventilation problem. Following the test tasks, the participants' satisfaction with the device interface was assessed by means of open and closed questions (Likert scale). The developed usability test was performed unsupervised online by doctors experienced in ventilating patients. To obtain a first impression regarding the applicability of the tool for usability testing, the users are asked questions about the suitability and comprehensibility of the tool and tool-specific anomalies in the collected data are analysed.

4. Discussion und Conclusion

The online survey tool SoSci Survey was extended to conduct a remote usability test using a prototype of a ventilator interface. The tool allows the integration of classic usability test content to a limited extent. The prototype must be available in a suitable format, which can be integrated via the methods provided. The prototype developed in advance could only be integrated as recorded videos. However, this was sufficient for evaluating the comprehensibility of the displayed content. The flexibility of the test person in terms of time and place goes hand in hand with restrictions in the design of the test. Functionalities of on-site tests such as interaction within a realistic context of use and detailed questions to clarify specific aspects are not possible. Due to these limitations, on-site tests with users cannot be replaced for the described use case. Nevertheless, this form of testing can provide initial results for rapid further development (2,4). The low response rate shows that intensive work is still required in the recruitment of test persons.

References

[1] ISO 62366-1:2015, Medical devices - Part 1: Application of usability engineering to medical devices.
[2] Spaeth J, Schweizer T, Schmutz A, Buerkle H, Schumann S. Comparative usability of modern anaesthesia ventilators: a human factors study. British journal of anaesthesia. 2017;119(5):1000–8.
[3] McLaughlin AC, DeLucia PR, Drews FA, Vaughn-Cooke M, Kumar A, Nesbitt RR, Cluff K. Evaluating Medical Devices Remotely: Current Methods and Potential Innovations. Human Factors. 2020;62(7):1041–60.
[4] Andreasen MS, Nielsen HV, Schrøder SO, Stage J. What Happened to Remote Usability Testing? An Empirical Study of Three Methods. Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '07). 2007;1405–14.

Challenges of Trustable AI and Added-Value on Health
B. Séroussi et al. (Eds.)

doi:10.3233/SHTI220591

Assessment of the Body Posture of Interventional Radiologists

Saskia LOHMANN[a], Gürkan ÖZTÜRK[a], Denise JUNGER[a], Rüdiger HOFFMANN[b] and Oliver BURGERT[a,1]
[a] *School of Informatics, Reutlingen University, Reutlingen, Germany*
[b] *Department of Radiology, University Hospital Tübingen, Tübingen, Germany*

Abstract. Physicians in interventional radiology are exposed to high physical stress. To avoid negative long-term effects resulting from unergonomic working conditions, we demonstrated the feasibility of a system that gives feedback about unergonomic situations arising during the intervention based on the Azure Kinect camera. The overall feasibility of the approach could be shown.

Keywords. Ergonomics, Human Factors, Interventional Radiology, Body Posture

1. Introduction

Physicians in interventional radiology are exposed to high physical stress. During interventions, they must spend long periods of time in static postures [1] while wearing heavy radiation protective clothing. Due to the required concentration during such interventions, the interventionist may not notice if he or she is working in an unergonomic body posture. This can lead to diseases such as disc herniations, which can limit the ability of those affected to perform their occupation [2]. Therefore, we aim on developing a system that can distinguish between ergonomic and unergonomic body postures. This information can be given to the interventionist to increase the awareness and maybe trigger a re-positioning of either patient, instruments, or interventional radiologist.

2. Methods

We did a use case analysis together with experienced interventional radiologists and derived 16 functional and non-functional requirements for such a system.

To describe a posture, a kinematic skeleton model can be used. We evaluated the following established methods for body posture analysis in working environments regarding their applicability to our project: DIN EN 1005-4 [4], PERA [5], work of Hellig [6], Ray [7], EAWS [8], and the work of Snook, NIOSH, RULA, REBA, MTM, and OWAS [9]. For the first iteration of the system, we implemented a decision table based on DIN EN 1005 combined with the criteria of the PERA system. In a second iteration, the RULA system was implemented as a second assessment option.

[1] Corresponding Author, Oliver Burgert, Faculty of Informatics, Reutlingen University, Alteburgstr. 150, 72762 Reutlingen, Germany; E-mail: oliver.burgert@reutlingen-university.de.

Due to the COVID-19 pandemic, pre-clinical tests were restricted to the research-OR at Reutlingen University. The testing clinician performed a simulated intervention wearing a lead protection (weight 10 kg) under a surgical gown.

3. Results

The system was successfully implemented (see Figure 1). The detection of the relevant joints was sufficient to demonstrate the general applicability of the approach. Nonetheless, we experienced many miss-detections, mainly caused by occlusions of the interventional radiologist or by the shape of the surgical gown.

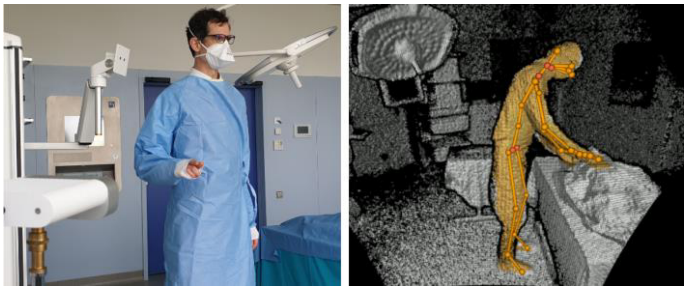


Figure 1. Setting in the research-OR (left), visualization of the skeleton model during the procedure (right).

4. Discussion and Conclusion

We were able to demonstrate the feasibility of body posture tracking in an interventional radiology setting for assessing the ergonomic situation of the interventional radiologist.

References

[1] Berguer R. Surgery and ergonomics. Arch Surg. 1999;134(9):1011–6. doi:10.1001/archsurg.134.9.1011.
[2] Cardenas-Trowers O, Kjellsson K, Hatch K. Ergonomics: making the OR a comfortable place. Int Urogynecol J. 2018;29(7):1065–6. doi:10.1007/s00192-018-3674-7.
[3] Zanuttigh P, Marin G, Dal Mutto C, Dominio F, Minto L, Cortelazzo GM. Time-of-Flight and Structured Light Depth Cameras: Technology and Applications. Cham, s.l.: Springer International Publishing; 2016. 355 p. doi:10.1007/978-3-319-30973-6.
[4] DIN EN 1005-4:2009-01, Sicherheit von Maschinen_- Menschliche körperliche Leistung_- Teil_4: Bewertung von Körperhaltungen und Bewegungen bei der Arbeit an Maschinen; Deutsche Fassung EN_1005-4:2005+A1:2008. Berlin: Beuth Verlag GmbH. ger.
[5] Chander DS, Cavatorta MP. An observational method for Postural Ergonomic Risk Assessment (PERA). International Journal of Industrial Ergonomics. 2017;57:32–41. doi:10.1016/j.ergon.2016.11.007.
[6] Hellig T. Musculoskeletal exertion under consideration of the interaction of working postures: RWTH Aachen University; 2019. ger. doi: 10.18154/RWTH-2019-10282.
[7] Ray SJ, Teizer J. Real-time construction worker posture analysis for ergonomics training. Advanced Engineering Informatics. 2012;26(2):439-55.
[8] Schaub K, Caragnano G, Britzke B, Bruder R. The European Assembly Worksheet. Theoretical Issues in Ergonomics Science. 2013;14(6):616–39. doi:10.1080/1463922X.2012.678283.
[9] Bullinger-Hoffmann AC, Mühlstedt J, editors. Homo Sapiens Digitalis - Virtuelle Ergonomie und digitale Menschmodelle. Wiesbaden, Heidelberg: Springer Vieweg; 2016. 406 p. ger.

Challenges of Trustable AI and Added-Value on Health
B. Séroussi et al. (Eds.)

doi:10.3233/SHTI220592

Diverse Recruitment Strategies Are Needed to Reduce Digital Divide: Results from a Workshop Addressing Digital Divide and Effects of Pandemic Restrictions

Meghan BRADWAY[a,1] and Eirik ÅRSAND[b,a]
[a] *Norwegian Centre for E-health Research, University Hospital, Tromsø, Norway*
[b] *Department of Computer Science, The Arctic University of Norway, Tromsø, Norway*

Abstract. Recruitment is a bottleneck for research - especially digital health studies. Studies often focus on those who are easy to reach or already engaged in their health, leaving those who are uninterested or un-engaged, as "un-reached". This contributes to the "digital divide". COVID-19 restrictions made recruitment more difficult. During a virtual workshop of our peers, we discussed recruitment of un-reached groups for digital health studies, especially during COVID-times. All agreed; we need to go where the un-reached are by collaborating with community-based services and organizations.

Keywords. Digital divide, un-reached, recruitment, COVID-19

1. Introduction

In digital health research, those who lack resources or interest to engage in their health, or are unable to use available technology and services, are often excluded. Research recruitment methods may be one reason that certain groups are "un-reached". This contributes to the "digital divide" [1]. COVID-19 pandemic restrictions further increased the digital divide and recruitment challenges.Even more of our educational, work and social activities are forced to be online [2]. Literature notes that research has followed this same trend; in response to social distancing restrictions, recruitment activities are moved online [3]. However, this solution still only reached those who are engaged in their health and/or have online access, leaving those who are not, un-reached. Therefore, we need to be more creative in our recruitment methods and channels in order to engage those who are un-reached in digital health research.

2. Methods

We arranged a 3-hour virtual workshop on November 19, 2021 to discuss ongoing challenges of, and possible solutions for, recruitment of those who are usually un-reached by digital health studies, especially during COVID times.

[1] Meghan Bradway, University Hospital of North Norway's Norwegian Centre for E-health Research, Tromsø, Norway, email: Meghan.Bradway@ehealthresearch.no

3. Results

First, seven participating researchers presented strategies used to adapt their recruitment to COVID times, e.g. using social media and posting QR-codes throughout the community. Next, participants discussed largely un-reached groups, limitations for recruitment and potential, targeted recruitment channels. Un-reached groups included those who: were technology or health-management averse, had substance abuse problems, did not have permanent homes and those with "alternative views" (e.g. conspiracy theorists). Challenges to recruitment included the fact that individuals may: not encounter online recruitment, lack funds to consistently pay for mobile phone and data plans, lack access to health and technology service, e.g. ability to charge mobile devices' batteries, lack personal security, or rely on misinformation that fuels distrust in society, politics and healthcare. The main challenge, of course, was limited access to in-person services due to social distancing policies. While some would require reduced restrictions, targeted recruitment channels included churches, groceries, libraries, shelters, and closed Facebook groups for a specific groups' unique interests. The common theme being that we needed to exert more effort when tailoring our recruitment and expand past our common strategies.

4. Discussion

Researchers' recruitment changes still follow the same trend as the public - moving activities online and using technologies to which not everybody has access. Common recruitment barriers highlighted challenges that many take for granted, e.g. being able to re-charge personal devices. All agreed that community organizations should be used for recruitment as they are more widely accessible. This is supported by the literature [4]. Even as pandemic restrictions are eased, we will continue to use the convenient online solutions that emerged in response to COVID. This will allow the digital divide and lack of participation from the un-reached to persist if not addressed.

5. Conclusion

COVID-19 has forced us all to rely on technology more than ever, yet not everyone has the access or desire needed to participate. We as researchers have the opportunity and duty to adapt - to "go to the people" to address the digital divide during COVID times.
We would like to thank Helse Nord for funding these activities (HNF1444-19).

References

[1]. Makri, A., Bridging the digital divide in health care. The Lancet Digital Health, 2019. 1(5): p. e204-e205.
[2] Lai, J. and N.O. Widmar, Revisiting the digital divide in the COVID‐19 era. Applied economic perspectives and policy, 2021. 43(1): p. 458-464.
[3] Mitchell, E.J., et al., It is unprecedented: trial management during the COVID-19 pandemic and beyond. Trials, 2020. 21(1): p. 784.
[4]. Bonevski, B., et al., Reaching the hard-to-reach: a systematic review of strategies for improving health and medical research with socially disadvantaged groups. BMC medical research methodology, 2014. 14(1): p. 1-29.

Challenges of Trustable AI and Added-Value on Health
B. Séroussi et al. (Eds.)

doi:10.3233/SHTI220593

Evaluating mHealth Design for People with Dementia: Preliminary Results

Thomas ENGELSMA[a,b,1], Laila R. RIPPIN[a,b], Monique W. JASPERS[a,b] and Linda W. PEUTE[a,b]
[a] *Center for Human Factors Engineering of Health Information Technology*
[b] *Department of Medical Informatics, Amsterdam Public Health research institute, Amsterdam UMC, location AMC, Amsterdam, The Netherlands*

Abstract. The availability of mHealth for people with dementia is increasing. Various mHealth design guidelines for this population have been proposed. In this study, we developed a binary checklist with evaluation statements to assess the implementation of twenty design suggestions in seven currently available mHealth apps for people with dementia. Between 17%-65% of the evaluation statements in the checklist were implemented in these apps. Not all statements were considered applicable for each assessed mHealth app, which resulted in dividing the criteria in two groups as either key evaluation statements or optional evaluation statements. In future work we want to augment this checklist to contribute to the future design of mHealth for people with dementia.

Keywords. Design guidelines, mHealth, Dementia, Evaluation

1. Introduction

A rising number of mobile health (mHealth) apps are available for people with dementia (PwD) and can be a valuable addition to, for example, activities of daily living. However, their design does not always match the needs and capabilities of PwD. To contribute to the development of mHealth design guidelines for PwD, a previous scoping review identified twenty, expert- and evidence based, dementia-related design suggestions aiming to improve mHealth usability for PwD [1]. However, little is known about the actual implementation of these design suggestions in current mHealth apps available to PwD. Therefore, this short paper presents an assessment of currently available mHealth apps for PwD on these previously identified design suggestions. Ultimately, we aim to develop mHealth design criteria for PwD.

2. Methods

The twenty design suggestions from the scoping review were first reformulated into evaluation statements. The following structured sentence was applied to ensure

[1] Corresponding Author, Thomas Engelsma, Amsterdam UMC, University of Amsterdam, Department of Medical Informatics, Center for Human Factors Engineering of Health Information Technology, Meibergdreef 9, Amsterdam, the Netherlands; E-mail: t.engelsma@amsterdamumc.nl

consistency: 'The app' followed by either 'allows', 'has', 'provides' or 'supports', and finalized with the design suggestion. Next, a binary scoring checklist was developed. Each mHealth app was assessed with the evaluation statements and assigned 1 (implemented), 0 (not implemented) or "N/A" (statement not applicable for the evaluated app). Afterwards, percentages were calculated by dividing the number of implemented statements by the total applicable statements. The iOS App store was searched with the following topics: dementia, memory, Alzheimer, screening, tracking, and cognition. Results were screened and selected if they complied with the following criteria: (1) app has a focus on PwD or those with cognitive impairments and (2) free to download without in-app purchases.

3. Results

Reformulating the design suggestions led to twenty-one evaluation statements related to cognitive barriers (n=6), physical ability barriers (n=1), perception barriers (n=6), and frame of mind barriers (n=8). Seven mHealth apps for PwD were selected for evaluation of which two aim to support cognitive training, four to facilitate daily activities, and one to improve awareness. Assessment of these apps showed that out of the twenty-one statements eleven were considered key evaluation statements applicable to all PwD apps studied and ten statements were optional and depended on the functionalities of a mHealth app. In total between 17 and 19 statements were applicable per included mHealth app leading to an assessment score that ranged between 17% - 65%. One statement scored "N/A" for all seven apps.

4. Discussion & Conclusion

Twenty out of the 21 evaluation statements were found to be applicable for the assessment of mHealth apps for people with dementia. In general, the evaluated apps scored poor on the implementation of available design suggestions. The cause may be that some design suggestions are customizable in the system settings of the mobile device that runs the app (e.g. screen brightness) rather than being implemented as an in-app setting. Further development of mHealth design criteria for PwD should be tuned to specific mHealth functionalities while taking into account system versus in-app settings.

References

[1] Engelsma T, Jaspers MWM, Peute LW. Considerate mHealth design for older adults with Alzheimer's disease and related dementias (ADRD): A scoping review on usability barriers and design suggestions. Int J Med Inform. 2021 Aug;152:104494.

Challenges of Trustable AI and Added-Value on Health
B. Séroussi et al. (Eds.)

doi:10.3233/SHTI220594

Semi-Structured Interviews to Evaluate a BCMA Implementation Trouble Areas

Juan Marcos DESCALZO[a,1], Melanie CASSARINO[a], María Florencia GRANDE RATTI[a], María SMITH[a] and Daniel LUNA[a]
[a]*Health Informatics Department, Hospital Italiano de Buenos Aires, Argentina*

Abstract. Errors in medication administration involve risks to patient safety. "Bar-Coding Medication Administration" is implemented to prevent these errors. Adoption by nurses is one of the main determinants of their effectiveness. The Hospital Italiano de Buenos Aires implemented BCMA 6 years ago, but its adoption rate still finds resistance in certain sectors. We conducted semi-structured interviews with nursing staff to explore the barriers to the use in low-usage wards and explore the current perceptions of nurses. While nurses recognised the safety and usefulness of the BCMA system, they reported many difficulties. The feedback obtained through this process was useful for the implementation team to plan future interventions, priorities and improvements on the system. The semi-structured interview methodology proved useful as a continuous monitoring strategy.

Keywords. Nursing Informatics, Interview, BCMA

1. Introduction

Errors in medication administration involve risks to patient safety and are one of the most common causes of harm [1]. Drug delivery systems using Bar-Coding Medication Administration (BCMA) are implemented in health institutions to prevent these errors [2]. Nursing staff adoption of BCMA is one of the main determinants of their effectiveness [3]. Nurses are key actors in error prevention and may be legally liable for errors and negligence during administration in Argentina. The medication administration process is one of nurses' main tasks. Monitoring, evaluation and follow-up of nursing work is recommended for successful implementation [4]. However, studies focusing on nurses' perceptions and workflow are not common [2,4].

The Hospital Italiano de Buenos Aires (HIBA) implemented BCMA in 2015-16, but its adoption rate still finds resistance in certain sectors. A semi-structured interview model was chosen to explore this issue. Allowing using a question guide but also enabling space for reciprocal conversation, emergence of topics not considered by the interviewers, behavior exploration, individual expression, and other advantages [5].

The present work aims to explore the barriers to the use of the implemented BCMA system in low-usage wards and explore the current perceptions of nurses.

[1] Juan Marcos Descalzo, Health Informatics Department, Hospital Italiano de Buenos Aires, Tte. Gral. Juan D. Perón 4190, C1199ABB Buenos Aires, Argentina; E-mail: juan.descalzo@hospitalitaliano.org.ar.

2. Methods

The Hospital Italiano is a high complexity Stage 7 HIMSS accredited medical center in the City of Buenos Aires. It employs more than 9000 people, including 1600 nurses.

The interviews were conducted to nurses from 4 selected wards with the lowest drug identification using BCMA (as of August 2019). In total 12 volunteer nurses participated, 3 from each ward. An ad hoc guide was used, containing questions and dimensions related to work and organizational aspects in professional practice, system undermining factors, mobile stations, application and current state of implementation. For the analysis, textual recording of the interviews was performed and reviewed.

3. Results

As negative aspects, nurses mentioned station size as impractical, short battery life and slow tablets. Infrastructure issues, such as low WiFi coverage were also mentioned. QR code readers tend to break and there are often delays in their replacement or repair. Some nurses expressed concern about infection risk in isolated patients.

On the positive side, nurses stressed that the application is easy to understand, regardless of training processes. Rapid support from the Help Desk was highlighted.

4. Conclusion

Semi-structured interviews with real system users allowed the implementation team to know first-hand their reality, identify pain points and plan improvement strategies and reinforce the positive aspects. Continuous monitoring with open communication is key to successful implementation. We believe that this work tells a valuable experience regarding a Nursing IT implementation in Argentina, a topic and setting underrepresented in the literature. In addition, they were triggers for other lines of work, such as incorporating new app functionalities and exploring mobile alternatives.

References

[1] Bates DW, Singh H. (2018). Two decades since to err is human: an assessment of progress and emerging priorities in patient safety. Health Affairs, 37(11), 1736-1743.

[2] Godshall M, Riehl M. (2018). Preventing medication errors in the information age. Nursing 2020, 48(9), 56-58.

[3] Taliercio V, Schachner B. Borbolla, D., Luna, D. R., Villalba, E., & de Quirós, F. G. B. (2014, January). The expectations of nurses about the implementation of a Bar Coded Medication Administration System: a qualitative study. In MIE (pp. 191-195).

[4] Voshall B, Piscotty R, Lawrence J, Targosz M. (2013). Barcode medication administration work-arounds: a systematic review and implications for nurse executives. JONA: The Journal of Nursing Administration, 43(10), 530-535.

[5] Kallio H, Pietilä AM, Johnson M, Kangasniemi M. (2016). Systematic methodological review: developing a framework for a qualitative semi-structured interview guide. Journal of advanced nursing, 72(12), 2954-2965.

Challenges of Trustable AI and Added-Value on Health
B. Séroussi et al. (Eds.)

doi:10.3233/SHTI220595

Observed High Adherence to Recombinant Human Growth Hormone Treatment Using a Multi-Component Approach to Improve Adherence in Individuals with Growth Disorders

Kate MORGAN[a,1], Paula VAN DOMMELEN[b], Ana-Rita BATISTA[a], Ekaterina KOLEDOVA[c]
[a] *Merck Serono Ltd., Feltham, UK, an affiliate of Merck KGaA*
[b] *The Netherlands Organization for Applied Scientific Research TNO, Leiden, The Netherlands*
[c] *Merck Healthcare KGaA, Darmstadt, Germany*

Abstract. We explored whether a multi-component approach – using a digital health device, the easypod™ auto-injector, the 'MySupport' patient support programme (PSP) and a Patient Activation Measure® (PAM®) – could improve adherence in patients receiving recombinant human growth hormone (r-hGH). A 13-item PAM was used to assess caregiver self-reported knowledge, resulting in two PAM scores for 88 patients at four UK hospitals after an average of 5.6 months. Most patients improved their PAM score by ≥1 level (43%) or maintained it (>-1 and <1; 21%). In parallel, 74% of patients maintained (-5 to +5%) or improved (≥5%) their adherence. Further studies are required to evaluate a multi-component approach to adherence in a larger population and for a longer duration.

Keywords. Auto-injector, growth hormone, patient activation measure, patient support programme

1. Introduction

To initiate catch-up growth and improve adult height, a long-term commitment from patients and caregivers to regular injections of recombinant human growth hormone (r-hGH; somatropin; Saizen®, Merck Healthcare KGaA, Darmstadt, Germany) is required. However, poor adherence to r-hGH therapy is common and leads to reduced efficacy and a multi-component, individualized approach is generally required to increase motivation to improve adherence.[1] Here, we briefly explore some of these components and describe how additional incorporation of a Patient Activation Measure® (PAM®) may facilitate caregiver engagement and, in turn, patient adherence.

[1] Corresponding Author, Kate Morgan, Patient Support UK & IRE, Merck Serono Ltd., Feltham, UK, an affiliate of Merck KGaA; E-mail: Kate.Morgan@merckgroup.com.

2. The easypod™ connect ecosystem and patient support programmes (PSPs)

The easypod™ auto-injector device automatically records and transmits the date, time and dose of r-hGH injected and makes these data accessible to healthcare professionals (HCPs) via the easypod™ connect ecosystem.[2] For patients receiving r-hGH treatment in the UK, Merck Serono Ltd. (an affiliate of Merck KGaA) provides a PSP called MySupport, which is run by Lloyds Pharmacy Clinical Homecare.[3] The programme includes initial nurse training based on individual patient needs, ongoing remote telephone and digital support from MySupport nurses, conducted at 7 days, 30 days and 3 months after the initial nurse training visit (England and Wales), communication via telephone, text message, e-mail, post and other channels, dispensing and delivery of medication and ancillary items and ongoing contact with HCPs.

3. Patient Activation Measure

We used a validated 13-item PAM (licensed by Insignia Health, Portland, OR, USA) to collect two separate PAM scores for caregivers of 265 (PAM 1) and 113 (PAM 2) patients receiving r-hGH at four UK hospitals. The response categories were '**Agree Strongly**', '**Agree**', '**Disagree**' and '**Disagree Strongly**'. The majority of caregivers agreed (strongly) on all items (range, 96−100%). A total of 88 patients (out of n=265) – all of whom were supported via the MySupport PSP – had two PAM scores and showed a mean change in total score of +1.0 point, after an average of 5.6 months between assessments. The majority of patients improved (by ≥1 level [43%]) or maintained (>-1 and <1 [21%]) their PAM score. Such findings are consistent with a high level of activation conferred by MySupport among caregivers. In parallel, 74% of patients maintained (-5 to +5%) or improved (≥5%) their adherence.

4. Discussion and Conclusions

Activation was already very high among caregivers at the start of the study. When incorporated with a PSP ecosystem, we observed an improvement in PAM of 1.0 point after an average of 5.6 months. These findings highlight that the ideal PSP ecosystem should include a multi-component approach with a digital health platform, nursing support and patient/caregiver feedback. Further longer-term studies in a larger population, investigating the effect of such multi-component approaches to improving adherence to r-hGH therapy, are required.

References

[1] Acerini CL, Wac K, Bang P, Lehwalder D. Optimizing patient management and adherence for children receiving growth hormone. Front Endocrinol (Lausanne). 2017;8:313.
[2] Koledova E, Stoyanov G, Ovbude L, Davies PSW. Adherence and long-term growth outcomes: results from the easypod™ connect observational study (ECOS) in paediatric patients with growth disorders. Endocr Connect. 2018;7(8):914-23.
[3] Lloyds Pharmacy Clinical Homecare. Available from: https://lpclinicalhomecare.co.uk/
[4] Patient Activation Measure® (PAM®). Insignia Health, Portland, OR, USA (https://www.insigniahealth.com/products/pam).

Challenges of Trustable AI and Added-Value on Health
B. Séroussi et al. (Eds.)

doi:10.3233/SHTI220596

Testing Medical Student Diagnostic Reasoning Using Clinical Data Visualizations

Lucille CHENG[a,1] and Yalini SENATHIRAJAH[a]
[a] *University of Pittsburgh School of Medicine, USA*

Abstract. This experiment aimed to (1) induce System-1-type diagnostic reasoning in medical students through the acquisition of cognitive user interface (UI) heuristics and (2) understand qualitatively how clinical data visualizations could enhance medical education. Third- and fourth-year medical students were presented patient cases through a novel electronic health record (EHR) design then asked to diagnose patients after being shown the cases either briefly and repeatedly (Group A) or twice over a longer period (Group B). Group A had higher accuracy than Group B. Findings support the possibility of inducing System-1 reasoning via UI heuristics and potential of integrating data visualizations in medical education.

Keywords. EHR, System-1-type diagnostic reasoning, medical education, clinical reasoning, clinical data visualization, graph

1. Introduction

Clinical decision-making can be described using the dual-process theory [1], which postulates that cognitive processes consist of the fast, intuitive System 1 commonly seen with seasoned physicians and the slow, deliberative System 2 more associated with learners like medical students [2][3]. However, Rosby et al. showed that inducing accurate System-1-type diagnostic reasoning in novices was possible by training students to use cognitive heuristics via rapid repeated exposures to clinical x-rays in contrast to longer, fewer exposures [4].

Here, we further explored that notion by (1) attempting to induce System-1-type diagnostic reasoning in inexperienced medical students through the acquisition of cognitive UI heuristics and (2) better understand the impact of clinical patient data visualizations on students' cognitive load and medical education.

2. Methods

Subjects were 15 third- and fourth-year medical students who had completed at least one clinical rotation, randomized to two matched groups based on past EHR experience. They were first trained with eight adapted patient cases with clinical data indicative of non-alcoholic fatty liver disease (NAFLD) and/or metabolic syndrome, on a novel EHR

[1] Lucille Cheng, University of Pittsburgh School of Medicine, 3550 Terrace St, Pittsburgh PA USA 15213; E-mail: Luc56@pitt.edu.

visualization (Figure 1). Each group was then shown four of eight cases repeatedly, but with Group A up to four times at 30 seconds/case and Group B only twice at 2 minutes/case. In a final test, they were shown all eight cases, asked to diagnose the patient (i.e., has metabolic syndrome and/or NAFLD) and elaborate on their experience. Our null hypothesis (N_0) was that exposure frequency has no impact on accuracy.

Figure 1. One of eight cases participants on the UI compared to the conventional EMR screen from Epic (right). All cases were displayed on the same UI based on the open-source visualization h-graph (h-graph.org).

3. Results

Group A scored slightly higher on average than Group B, with a mean percentage correct of 0.76 (95% [0.68, 0.84]) vs 0.69 (95% [0.58, 0.80]). We failed to reject the null hypothesis (p-value = 0.40). 73% of participants (n=11) rated the new version on par or higher than existing EHRs (3+/5). Its ease of use and intuitiveness was rated similarly high (mean score = 3.73/5 and 4.2/5, respectively). Students also described "pattern-recognition" strategies consistent with System-1 decision-making and Rosby et. al [3]. These include: (1) layout consistency aiding search, (2) trends between lab value displays and subsequent diagnosis, and (3) visualization-specific features like color-coordination.

4. Discussion and Conclusion

Though not new, this type of clinician-side data visualization is rarely utilized and has promising implications for medical education enhancements. Students diagnosed patients more accurately after short, repeated exposure to the data visualization interface, implying the possibility of inducing Type-1 diagnostics. Using such techniques during care delivery could reduce cognitive burden and allow even novices to diagnose quickly and correctly. This area is ripe for innovation and further research for medical education.

References

[1] Evans JS. Dual-processing accounts of reasoning, judgment, and social cognition. Annu Rev Psychol. 2008;59:255–278.
[2] Djulbegovic B, Hozo I, Beckstead J, Tsalatsanis A, Pauker SG. Dual processing model of medical decision-making. BMC Med Inform Decis Mak. 2012 Sep;12:94.
[3] Rosby LV, Rotgans JI, Tan G, Low-Beer N, Mamede S, Zwaan L et al. Inducing System-1-type diagnostic reasoning in second-year medical students within 15 minutes. Medical Teach. 2018 Oct;40(10):1030-1035.

Challenges of Trustable AI and Added-Value on Health
B. Séroussi et al. (Eds.)

doi:10.3233/SHTI220597

Artificial Intelligence in Undergraduate Medical Education

Lina MOSCH[a,b] , Louis AGHA-MIR-SALIM[a], Merve M. SARICA[a], Felix BALZER[a], Akira-Sebastian PONCETTE[a,b1]
[a]*Institute of Medical Informatics, Charité – Universitätsmedizin Berlin, Germany*
[b]*Department of Anesthesiology and Intensive Care Medicine, Charité – Universitätsmedizin Berlin, Germany*

Abstract. AI will take on an increasingly important role in medicine. Therefore, AI competencies should be taught in medical school. We investigated the inventory of AI-related courses at German medical schools. The majority of faculty offer courses on AI, but mainly at the elective and introductory levels. Regarding the topic of AI, there is a gap in German medical education that should be closed.

Keywords. AI, medical education, qualification, competencies

1. Introduction

The increasing use of artificial intelligence (AI) in the medical field requires healthcare professionals who are able to work with the growing number of different medical AI systems, including the evaluation of their output and understanding of their limitations [1,2]. Medical students in Europe do not feel prepared for working in a digitalized healthcare system, and there is a need to include AI- and digital competencies more in medical education [3]. However, there is limited evidence on the status quo of educational programs in AI for undergraduate medical students in Europe.

We aimed to assess learning opportunities on AI at German medical schools.

2. Methods

In April 2021, following IRB approval [4], all 39 German medical faculties were consulted with an online questionnaire on AI-related learning opportunities. In addition, information on this was researched in publicly accessible sources (e.g. faculty websites, module handbooks etc.). Finally, all deaneries were asked to validate the compiled information; 29 of 39 institutions validated our information or completed the questionnaire themselves.

[1] Corresponding Author, Akira-Sebastian Poncette, Institute of Medical Informatics, Charité – Universitätsmedizin Berlin, Charitéplatz 1, 10117 Berlin, Germany; E-mail: akira-sebastian.poncette@charite.de

3. Results

The majority (28/39, 71.8%) of medical schools in Germany offer AI-related courses to students, mostly as elective courses or extracurricular activities. About three-quarters (31/39, 79.5%) of institutions are planning (more) AI-related learning opportunities.

Most common are courses that provide an overview and introduction to the application areas of AI. At nine institutions, such courses are anchored in the core curriculum. Almost all (25/28, 89.3%) institutions with AI courses consider ethical, legal, and social issues of AI, as well as the special needs of communicating with patients using AI applications.

15 of 38 (39.5%) faculties reported having implemented courses on AI for the first time in 2020 or 2021.

4. Discussion

AI-related learning content is mainly present in voluntary courses, but hardly in the compulsory curriculum of medical studies. However, we see dynamic development in this area, with the majority of medical schools planning to introduce AI courses and a large proportion of learning opportunities introduced in the last two years.

5. Conclusions

Basic competencies in the field of AI are essential to enable physicians to supervise AI systems but are not taught comprehensively in medical training in Germany. They should become an integral part of the mandatory curriculum at medical school.

References

[1] Topol E. Preparing the healthcare workforce to deliver the digital future [Internet]. 2019. Available from: https://topol.hee.nhs.uk/wp-content/uploads/HEE-Topol-Review-2019.pdf.

[2] Benjamens S, Dhunnoo P, Meskó B. The state of artificial intelligence-based FDA-approved medical devices and algorithms: an online database. npj Digit Med 2020 Sep 11;3(1):1–8. [doi: 10.1038/s41746- 020-00324-0]

[3] Machleid F, Kaczmarczyk R, Johann D, Balčiūnas J, Atienza-Carbonell B, Maltzahn F et al. Perceptions of Digital Health Education Among European Medical Students: Mixed Methods Survey. Journal of Medical Internet Research 2020;22(8):e19827. [doi: 10.2196/19827]

[4] Ethikausschuss am Campus Charité - Mitte, Vote: EA1/236/19, Chairperson: Prof. Dr. R. Morgenstern

Challenges of Trustable AI and Added-Value on Health
B. Séroussi et al. (Eds.)

doi:10.3233/SHTI220598

Jupyter Notebooks for Introducing Data Science to Novice Users

Mathilde FRUCHART[a,1] , Benjamin GUINHOUYA[a,b], Sylvia PELAYO[a,b], Christian VILHELM[a,b] and Antoine LAMER[a,b]
[a]*Univ. Lille, Faculté Ingénierie et Management de la Santé, F-59000, Lille, France.*
[b]*Univ. Lille, CHU Lille, ULR 2694 - METRICS: Évaluation des Technologies de santé et des Pratiques médicales, F-59000 Lille, France.*

Abstract. Data science is a bridge discipline involving computer science, statistics, and knowledge of the health field. We developed a Jupyter Notebook to enable novice users to easily and autonomously analyze data from social networks. We conducted an experimentation with non-programmer students. They had to adapt a R Notebook and complete 14 questions and to perform descriptive analyses. The average rate of correct answers was 90.7. Jupyter Notebook enabled novice users to easily and autonomously analyze data from Twitter.

Keywords. Data Science; Education; Social media; Programming

1. Introduction

Data science may seem to be restricted to expert users because it requires the mastery of many technologies and knowledges [1]. Analyses are generally programmed from scripts consisting of a series of instructions, accompanied by comments, and developed with an integrated development environment (IDE) or in command line. Notebooks are another approach to programming and sharing results and were designed to improve the reproducibility of the studies [2]. The final notebook provides a full and comprehensive analysis report. We developed a R notebook to enable novice users to easily and autonomously analyze data from social networks. We assessed their ability to use it and captured their feedback on the value of such a tool for their daily practice.

2. Methods

Jupyter Notebook is an interactive computational laboratory notebook, which can work with code in many different programming languages such as Python, Java, R, or Julia [3]. Jupyter Notebook allows for the smooth integration of code and narrative text (in Markdown syntax) into a single document that can be executed and edited on the fly. We experimented the Jupyter Notebook with students from two master's degree formations not specialized in Data Science and without programming skills.

[1] Corresponding Author, Mathilde Fruchart, ULR 2694, Lille University, 2 place de Verdun, F-59000, Lille, France; E-mail: mathilde.fruchart@univ-lille.fr.

Wedistributed a R Notebook with datasets from Twitter. Students had to execute the Notebook and to adapt the script in order to obtain additional results.

We assessed the ability of the students to use the Jupyter Notebook in checking if they were able to perform the 14 tasks [4]: (i) instructions in relation to the notebook (modify and complete cells, execute cells, create new cells, export in pdf); (ii) R code 1st level (from an existing code, compute a new variable, change a color in a graphic, change a variable in a graphic); (iii) R code 2st level (find a instruction already implemented and modify it to produce a new result).

3. Results

We have delivered 20 health topics to 44 students. For each topic, we extracted 1000 tweets. Average (SD) rate of correct answers was 90.7 (13.3) for all the questions. Questions with the poorest scores were about (i) the modification of the required dataframe and the variable for the generation of a barplot, (ii) the updating of an instruction to display mentions instead of hastags and the adaptation of the title of the graphic, (iii) the modification of cells in Markdown format or (iv) the creation of two new cells for code or Mardown. The main difficulties that emerged were ensuring that users executed all the code cells, and that they had set up their cells in Markdown or code according to their needs. On the contrary, we did not report any major error on syntax problems which we usually encounter when learning a new language.

4. Discussion and conclusion

We developed a R notebook to enable novice users to easily and autonomously analyze data from Twitter. We tried it out with a class of 44 students without programming skills. Compared to Rmarkdown, which also integrate code and narrative text, Jupyter Notebook prevents the need to work locally and to have to install the necessary libraries himself [5]. IDE are rather designed for more advenced developers because they provide interesting features, such as code autocompletion, and displays the environment with all variables. However they could be more complex to handle for novice users than Jupyter notebook. Improvements would be to not display warnings and messages to provide training on manipulation of the notebook.

References

[1] Parker MS, Burgess AE, Bourne PE, Ten simple rules for starting (and sustaining) an academic data science initiative, PLoS Comput. Biol., vol. 17, no 2, p. e1008628, févr. 2021.
[2] Kluyver T et al. Jupyter Notebooks – a publishing format for reproducible computational workflows, Position. Power Acad. Publ. Play. Agents Agendas, p. 87-90, 2016.
[3] Project Jupyter. https://jupyter.org.
[4] Mathilde Fruchart / Twitter analysis for novice users, GitLab. https://gitlab.com/mathilde.frchrt/twitter-analysis-for-novice-users.
[5] Baumer B, Cetinkaya-Rundel M, Bray A, Loi L, Horton NJ. R Markdown: Integrating A Reproducible Analysis Tool into Introductory Statistics, ArXiv14021894 Stat, févr. 2014.

Challenges of Trustable AI and Added-Value on Health
B. Séroussi et al. (Eds.)

doi:10.3233/SHTI220599

Digital Health Education: Determining Competences and Piloting Innovative Study Course

Georgi CHALTIKYAN[a], Fara Aninha FERNANDES[a,1], Jishen PFEIFFER[a]
[a] *Deggendorf Institute of Technology, European Campus Rottal-Inn, Germany*

Abstract. Digital technology for health services plays a critical role in the delivery of health services. In order to move towards universal healthcare, improvement of patient outcomes and better health, one must make use of the advantages of Digital Health tools and recognition of the role of the health ICT worker. Therefore, it is necessary to have a workforce that is competent to use these tools. Uniquely positioned at the intersection of healthcare and information technology, the domain of Digital Health builds on a variety of disciplines termed biomedical and health informatics, and other allied fields. With the increasing need to have a knowledgeable, skilled and competent workforce, it is necessary to concentrate efforts towards the provision of education modules in Digital Health. While continuing medical education, certificate courses and other similar courses attempt to bridge the gap in the delivery of Digital Health education, it is also paramount to establish dedicated and standalone courses. Streamlining approaches to Digital Health Education across disciplinary, cultural and national boundaries, is key to address the challenges of firmly embedding Digital Health courses in the fabric of university education. In the effort to provide the necessary knowledge, skills and competencies (KSCs) to the current health ICT worker, the Deggendorf Institute of Technology, European Campus Rottal-Inn (DIT-ECRI) is in the process of piloting a virtual course in Global Digital Health. With the ability to provide core competencies in Digital heath, this virtual course is a step towards advancing Global Digital Health Education.

Keywords. health workforce, digital health, information science, distance education, patient education

1. Introduction

1.1 The Scope of Digital Health

The move towards digital transformation, widely called the fourth industrial revolution presents a tectonic shift and the most important transition in the history of humankind. Digital Health technology offers a vital means to revolutionize the way national health systems are maintained. Supporting these initiatives is the necessity to educate professionals in the field of Digital Health. Such an approach is imminent so as to fully utilize the advantages and competencies of Digital Health, to prepare health professionals to work in a Digital Health environment and lastly, to empower the citizen.

[1] Corresponding Author, European Campus Rottal-Inn, Pfarrkirchen 84347 Germany; E-mail: fara.fds@gmail.com.

2. Innovative Virtual Course 'Global Digital Health' (IVC-GDH)

DIT-ECRI is currently heading a project to develop and pilot an innovative virtual student-oriented approach for studying the Digital Health environment in health systems in various countries. The goal of this project is to pilot a comprehensive innovative online (virtual) course of instruction in Global Digital Health, policies and practices in different countries and regions.

The course will be structured around the entire scope of the Digital Health domain. Students will be introduced to Digital Health technologies and systems such as digital health records, health information systems, interoperability, telehealth, mobile health, digital therapeutics, digital image analysis, and health analytics. As the scope of Digital Health grows, students will additionally learn the concepts of artificial intelligence, biomarkers, omics and precision medicine.

The core teaching methodology will be activities with students in studying health systems globally and the impact of Digital Health technologies through case studies and *Countrynalysis*™ (country profiles). Students will be encouraged to conduct an in-depth review of a country's health system and the Digital Health landscape [1].

A needs assessment survey was carried out to determine the knowledge, skills and competences of healthcare and IT professionals, medical and health educators, and all those with a career, knowledge or interest in the field of Digital Health [2]. Almost half of the 77 respondents from 26 countries reported that they were very familiar with the field of Digital health and its definition by the WHO Global Strategy on Digital Health 2020-2024. Around 67% of the respondents agreed that there is a need for a standalone interprofessional course in Digital Health. When asked about which topics or elements should be included in such a course, the respondents noted that ethics training, interdisciplinary training, training on research methods, hands-on IT training and project management training were most important.

3. Conclusion

Building a workforce with knowledge, competency, and skills for the adoption, initiation, and operation of digital technologies is essential for the full utilization of such technology in healthcare. Education and training in the core competencies are fundamental for the health worker and the health ICT worker. At DIT-ECRI, efforts in this direction are furthered through the Master of Digital Health course and the IVC-GDH course. In addition, it is necessary to establish a framework of a distance learning Digital Health education and training program to help educators around the world to improve, or further develop education in Digital Health.

References

[1] Chalitkyan G, Fernandes FA. Digital Health Education and Development of CONEDIG (Consortium of Educational Institutions in Digital Health). Workshop presented at EFMI MIE; 2021 May 29-31; Athens, Greece.

[2] Survey "Needs Assessment for Digital Health Education" [Internet]. Available from: https://survey.th-deg.de/index.php/167215?lang=en.

Section VII

Bioinformatics; Natural Language Processing

This page intentionally left blank

Challenges of Trustable AI and Added-Value on Health
B. Séroussi et al. (Eds.)

doi:10.3233/SHTI220601

Augmenting THerapeutic Effectiveness Through Novel Analytics (ATHENA) - A Public and Private Partnership Project Funded by the Flemish Government (VLAIO)

Ploingarm PETSOPHONSAKUL[a,1], Ashkan PIRMANI[b], Edward DE BROUWER[b], Murat AKAND[c], Wouter BOTERMANS[a], Frank VAN DER AA[c], Joris Robert VERMEESCH[d], Fritz OFFNER[e], Roel WUYTS[f], Yves MOREAU[b], Ingrid MAES[g], Ines BLOCKX[a], Patricia VAN ROMPUY[h], Martine LEWI[a] and Bart VANNIEUWENHUYSE[h]

[a] *Janssen R&D/Clinical Innovation, Belgium*
[b] *ESAT-STADIUS, KU Leuven, Belgium*
c *Dept. of Urology KU, Belgium*
[d] *Center for Human Genetics, UZ Leuven, Belgium*
[e] *Leuven, Dept. of hematology University of Ghent, Belgium*
f IMEC, Belgium
[g] *Inovigate, Belgium*
[h] *Janssen BeNeLux, Belgium*

Abstract. The complexity and heterogeneity of cancers leads to variable responses of patients to treatments and interventions. Developing models that accurately predict patient's care pathways using prognostic and predictive biomarkers is increasingly important in both clinical practice and scientific research. The main objective of the ATHENA project is to: (1) accelerate data driven precision medicine for two use cases – bladder cancer and multiple myeloma, (2) apply distributed and privacy-preserving analytical methods/ algorithms to stratify patients (decision support), (3) help healthcare professionals deliver earlier and better targeted treatments, and (4) explore care pathway automations and improve outcomes for each patient. Challenges associated with data sharing and integration will be addressed and an appropriate federated data ecosystem will be created, enabling an interoperable foundation for data exchange, analysis and interpretation. By combining multidisciplinary expertise and tackling knowledge gaps in ATHENA, we propose a novel federated privacy preserving platform for oncology research.

Keywords. Precision medicine, Federated platform, Machine learning, Distributed analytics, Data science

[1] Corresponding Author, Ploingarm Petsophonsakul; E-mail: ppetsoph@its.jnj.com.

1. Introduction

Pharmaceutical compounds and/or other therapeutic interventions may not show the same level of efficacy in all patients, leading to the term "imprecision medicine". Any approach that can enhance the effectiveness of therapy should be welcomed. There is a strong belief in the benefits of enriching available clinical patient information (non-omics data) with omics data which refers to large biomarker data sets characterizing biological features such as genomics, transcriptomics, proteomics etc. [1]. Omics data is considered as supporting data to gain a deeper understanding of the complex multifactorial causes and the natural evolution of a disease; as well as defining patient characteristics that can predict treatment success [2]. Such knowledge is needed for a wide range of diseases, most notably in oncology. Thus, improved disease insight in combination with predictive analytics form the basis of Personalized Medicine.

Project ATHENA (Augmenting THerapeutic Effectiveness through Novel Analytics), aims at gaining a deeper understanding on the various challenges of creating this combined omics and non-omics data approach for researching two cancer types: one solid tumor (bladder cancer) and one hematological cancer (multiple myeloma). Bladder cancer is selected because of its challenges in risk stratification, and multiple myeloma because it is regarded as one of the most complex cancer types exacerbated by a multitude of treatment options [3]. Improvements in targeted treatment will require the identification of relevant, actionable biomarkers to either risk stratify or adapt treatments. Identification of such markers requires analyses of data across the entire pheno-genotypic (omics and non-omics) set of characteristics and across a fully longitudinal follow-up of an entire patient cohort. This can create considerable privacy challenges. To overcome these, we will take a federated and privacy preserving analytics approach whereby data stays local under the control of the original data custodians.

Current state-of-the-art in Omics/Non-omics integration and Federated learning

Omics data integration has been addressed in recent years [4, 5]. However, only a few of them resulted in omics-based algorithms with sufficient predictive ability to be implemented into clinics or public health domains [2,6]. The relatively poor predictive ability of genomic data may be explained by the difficulty to analyse and extract relevant information from the omics data and by the large variation of health-related traits explained by non-omics data [7]. Therefore, it is crucial to integrate omics and non-omics data in the same models [8]. There is strong demand for federated systems that enable joint analysis efforts across multiple partners holding sensitive or competitively valuable data. Several projects aimed to address data privacy including Machine Learning Ledger Orchestration for Drug Discovery (MELLODDY), where machine learning is used to accelerate drug discovery while ensuring privacy preservation of both the data and the models through federated learning [9].

2. Concrete Objectives and Criteria

The main objective of the present project is to generate new knowledge to create a federated privacy preserving machine learning platform that can execute the newly researched machine learning algorithms according to medical case requirements.

2.1. The supporting info/infrastructure will require research on and integration of the following data pipelines:

- Non-omics data processing pipeline for clinical and patient recorded data.
- Non-omics imaging data processing pipeline with algorithms for the standardization and privacy preservation of imaging data.
- Standardized high performance pipelines for omics data processing.
- High performance somatic variant calling pipelines.
- Systems genetics pipeline for feature extraction from omics data.

A modular technical design will be used in support of data gathering and distributed analytics. Data pipelines (clinical, image and omics) and the output of each pipeline comes together in an integrated data warehouse for each of the participating institutions. These data warehouses form the basis for the exploration via distributed machine learning.

2.2. The performance of the platform should enable optimum data flow and linkage with proper user management incorporating appropriate user authentication and authorisation and thus be robust once scaled up.

2.3. Explore, research, and develop

- Novel capabilities for data from real-world patient trajectories.
- Federated and privacy-preserving implementations of these techniques that let consortium partners protect the privacy of patient data and maintain full control over the processing of these data at all times.
- Appropriate algorithms to ensure standardization and privacy protection of imaging data for the creation of the non-omics visualization pipeline.

2.4. Accelerate data driven precision medicine for two use cases – bladder cancer and multiple myeloma.

Bladder cancer: To develop a retro- and prospective longitudinal dataset of the entire population treated for non-muscle-invasive bladder cancer (NMIBC) including phenotype and genotype information, stored on local data sources and capable of generic data extraction to address NMIBC key scientific questions. **Multiple myeloma:** To develop a prospective longitudinal dataset of the entire population treated for multiple myeloma including phenotype and genotype information, stored on local data sources and capable of generic data extraction to address key scientific questions.

2.5. Define a governance framework (incl. legal, ethical, data privacy aspects) and conduct project management to successfully accomplish the project goals.

3. Approach

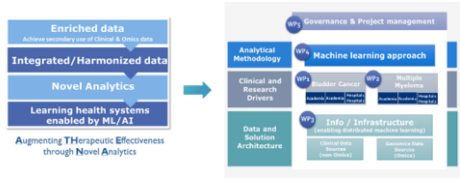


Figure 1. Structural approaches of ATHENA

The research will tackle the fundamental challenges associated with integrating omics and non-omics data by means of investigating and creating the appropriate data ecosystem that will enable an interoperable foundation for data exchange and capability. This will serve as the underlying info/infrastructure to facilitate the investigation and creation of a federated and privacy-preserving machine learning platform. A schematic overview of the ATHENA research project with different work packages (WPs) is provided in Figure. 1.

4. Deliverables (year 1)

1. Standardization of data catalogue and harmonization of NMIBC and multiple myeloma data
2. Independent Privacy/Security audit of local instance of Feder8 platform, a federated data network solution, and implementation of the central and local Feder8 component
3. High-level architecture blueprint for data pipelines integration
4. Approvals of technical and retrospective bladder cancer protocol

5. Impact and Expected Outcome

Resulting machine learning methodologies and the clinical outcomes of the project represent new knowledge that is set to advance the current state of the art in their respective fields. **Industry impact:** New knowledge on infrastructure, pipelines, warehouse data management, integration, security, standardization and privacy algorithms (analytics) etc., offering valuable economic impact for industry partners and opportunity for future development. **Independent research impact:** Proof of principle on new application-driven machine learning methodologies. Clinical outcomes offering new insights into bladder cancer and multiple myeloma for improved patient treatment. ATHENA is the new omics and non-omics approach to personalized medicine. **Clinical impact:** Insights in disease mechanisms and impact of different progression, optimal care and treatment pathways and patient risk-stratification for bladder cancer and multiple myeloma will be obtained by applying distributed analytics.

Within the ATHENA framework, a federated privacy preserving machine learning platform that can execute machine learning algorithms according to specific medical case requirements will be developed. This platform allows multiple hospitals to collaborate and build a common machine learning model without directly sharing sensitive data

(Figure. 2) Each institution remains in full control of its own data and resources. The platform offers privacy by design and complies with General Data Protection Regulation (GDPR) [10]. Expected project outcomes include (1) Removed barriers in ethics, consent and data governance, data quality and interoperability and affordability (2) Strengthened research partnership between academia and industry (3) Established regulatory frameworks and policy, as a basis for value-based healthcare.

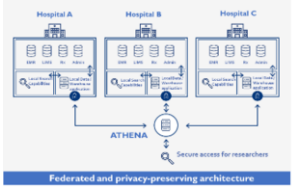


Figure 2. ATHENA privacy-preserving machine learning platform

6. Conclusions

ATHENA brings together a unique, multidisciplinary and complementary partnership of oncologists, experts in IT architecture, data science, high performance cloud computing, genomics and medical affairs. ATHENA utilizes leading-edge technology to integrate and enrich patient level data and analytics. The project leverages machine learning to generate insights from electronic health record (EHR), genomics and medical imaging data, thus creating a federated privacy-preserving machine learning platform that will accelerate data driven precision medicine and provide solutions to ethics, consent and data governance, data quality, interoperability and sustainability.

Acknowledgment

ATHENA is a public private partnership project funded by Flanders Innovation & Entrepreneurship (VLAIO). Project number: HBC.2019.2528

References

[1] Letai A. Functional precision cancer medicine-moving beyond pure genomics. Nat Med. 2017 Sep 8;23(9):1028-1035. doi: 10.1038/nm.4389. PMID: 28886003.
[2] Brandão M, Pondé N, Piccart-Gebhart M. Mammaprint™: a comprehensive review. Future Oncol. 2019 Jan;15(2):207-224. doi: 10.2217/fon-2018-0221. Epub 2018 Aug 29. PMID: 30156427.
[3] Lohr JG, et al. Widespread genetic heterogeneity in multiple myeloma: implications for targeted therapy. Cancer Cell. 2014 Jan 13;25(1):91-101.
[4] Hasin Y, Seldin M, Lusis A. Multi-omics approaches to disease. Genome Biol. 2017 May 5;18(1):83.
[5] Ritchie MD, Holzinger ER, Li R, Pendergrass SA, Kim D. Methods of integrating data to uncover genotype-phenotype interactions. Nat Rev Genet. 2015 Feb;16(2):85-97.
[6] Wallden F, et al. Development of the molecular diagnostic (MDx) DLBCL Lymphoma Subtyping Test (LST) on the nCounter Analysis System. J. Clin. Oncol. 2015;33.
[7] Lichtenstein P, et al. Environmental and heritable factors in the causation of cancer--analyses of cohorts of twins from Sweden, Denmark, and Finland. N Engl J Med. 2000 Jul 13;343(2):78-85.
[8] Gerstung M, et al. Precision oncology for acute myeloid leukemia using a knowledge bank approach. Nat Genet. 2017 Mar;49(3):332-340.
[9] https://www.melloddy.eu/ visited 16/03/2022.
[10] https://ec.europa.eu/info/law/law-topic/data-protection/data-protection-eu_en visited 16/03/2022.

Challenges of Trustable AI and Added-Value on Health
B. Séroussi et al. (Eds.)

doi:10.3233/SHTI220602

viroCapt: A Bioinformatics Pipeline for Identifying Viral Insertion in Human Host Genome

Maxime WACK[a,b,1], David VEYER[c,d], Camille PENEAU[e,f], Sonia LAMEIRAS[g], William DIGAN[a,b], Alain NICOLAS[g], Jessica ZUCMAN-ROSSI[e,f], Sandrine IMBEAUD[f], Anita BURGUN[a,b,h], Hélène PÉRÉ[c,d] and Bastien RANCE[a,b,h]

[a] *INSERM, UMRS 1138, Centre de Recherche des Cordeliers. Université de Paris, France*
[b] *Département d'Informatique Médicale, HEGP, AP-HP, France*
[c] *INSERM U970, PARCC, HEGP, Faculté de Médecine, Université de Paris, France*
[d] *Service de Microbiologie, HEGP, AP-HP, France*
[e] *Centre de Recherche des Cordeliers, INSERM, Université de Paris, France*
[f] *Functional Genomics of Solid Tumors laboratory, Labex OncoImmunology, Paris, France*
[g] *ICGex NGS platform, Institut Curie, PSL Research University, Paris, France*
[h] *Faculté de Médecine, Université de Paris, France*

Abstract. *Introduction*. The implication of viruses in human cancers, as well as the emergence of next generation sequencing has permitted to investigate further their role and pathophysiology in the development of this disease. One such mechanism is the integration of portions of viral genomes in the human genome, as well as the specific action of viral oncogenes.inding integration sites and preserved oncogenes is still relying on heavy manual intervention. *Methods*. We developed an analysis and interpretation pipeline to determine viral insertions. Using data from directed viral capture, the pipeline conducts a crude genotyping phase to select reference viral genomes, identifies chimeric reads, extracts the putative human sequences to locate in the human reference genome, scores and ranks candidate junctions, and exports tabular and visual results. *Results*. We leverage common bioinformatics tools (bowtie2, samtools, blat), and a dedicated filtering and ranking algorithm, implemented in R, to infer candidate junctions and insertions. Static results (tables, figures) are produced, as well as an interactive interpretation tool developed as a shiny web app. *Discussion*. We validated this pipeline against published results of HPV, HBV, and AAV2 insertions and show good information retrieval.

Keywords. bioinformatics, HPV, HBV, AAV2, virus, integration, cancer

[1] Corresponding Author, Maxime Wack, Département d'Informatique Médicale, HEGP, AP-HP, Paris, France; E-mail: maxime.wack@aphp.fr

1. Introduction

High-throughput data provided by next generation sequencing (NGS) have revolutionized the medical care of cancer patients. Together with treatment responses, the fine characterization of mutations in cancer cells has permitted the realization of precision medicine, and can now lead to personalized treatments in cancers. [1]

Recently, virus-induced cancers have received a great deal of attention. [2] Many viruses are implicated in the development of cancers, including HBV for hepatocellular carcinoma, [3] HPV for cervix, anal, and oropharyngeal squamous cell carcinoma, [4] EBV for nasopharynx carcinoma and Burkitt and Hodgkin lymphomaes, HTLV-1 for adult T-cell lymphoma, HIV and HHV8 for kaposi sarcoma.

The exact physiological mechanisms explaining the carcinogenic role of these viruses are not still clearly known, but recent advances in viral capture techniques followed by next generation sequencing have helped uncovering mechanisms underlying this carcinogenic role. [5,6]

Due to its high-throughput nature, NGS produces large amounts of data for which the bottleneck now lies in the interpretation of the results.We introduce `viroCapt`, a software package designed to help researcher in the analysis and interpretation of such data.

`viroCapt` manages an analytic and interpretation pipeline, as well as a visualization tool to assist in reading and interpreting the results, with a ranking method using quality and interpretability criteria.

2. Methods

The pipeline uses short read sequencing data from NGS viral capture, leveraging common bioinformatics tools (bowtie2 for viral alignment, [7] samtools for janitorial purposes, [8] and blat for human alignment [9]), and public resources (viral and human reference genomes) to produce candidate insertion sites in the human genome and splicing sites on the viral genome. Those results are ranked using custom-developed filters to assess their quality and interpretability.

The pipeline goes through the following steps (Figure 1): Quality Control, crude genotyping, local alignment to the candidate viruses, and alignment of the partially mapped reads to human, in a method described as Strategy A-II in Chen *et al.* 2019 [10]. Furthermore, we apply dedicated filters to tag specific human alignment (H), concordance of results between multiple reads (C), and presence of breakpoints on the same chromosome (T).

After the execution of the pipeline, the results can be browsed through a web app (Figure 2) showing the sequencing profile with the candidate insertions as an overlay, as well as sortable and filterable tabular view of all the results. Multiple options and filters (chromosome, quality, number of reads, length of the human sequence found) can be used to explore and narrow the results.

We validated the pipeline output on HeLa cell samples with an HPV integration and samples from hepato-cellular carcinomas with HBV and AAV2 integrations (data available upon request). [11,12,13]

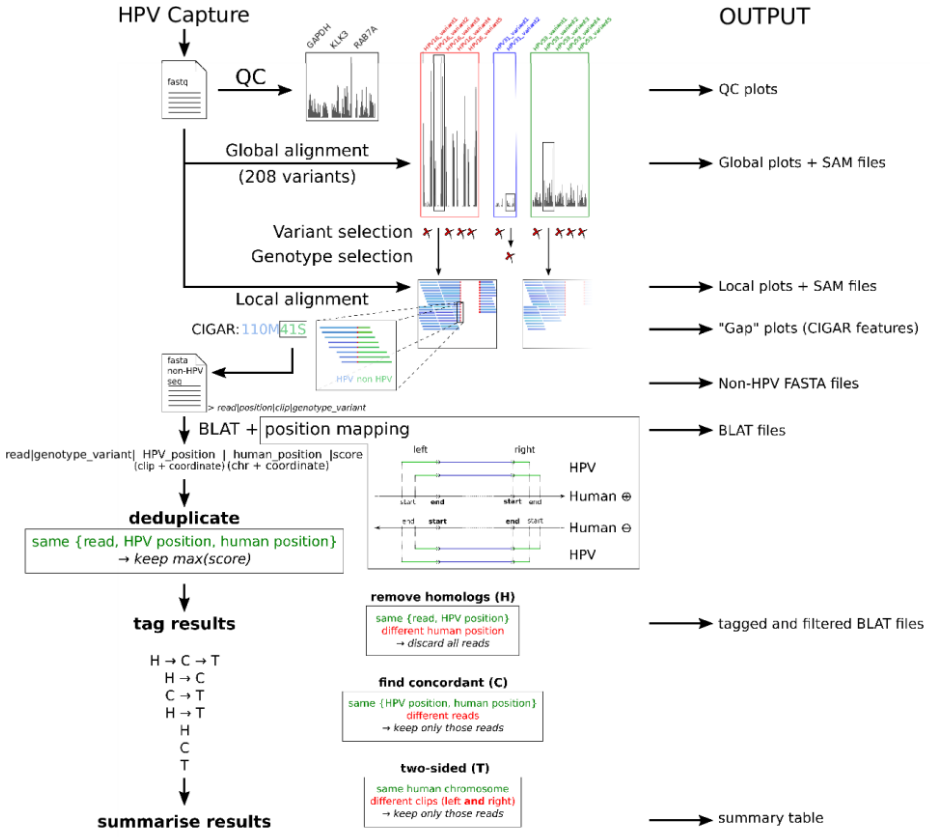


Figure 1. Overall organisation of the pipeline.

3. Results

The pipeline is implemented as a makefile orchestrating multiple bioinformatics tools, as well as R functions to process the intermediary data. Everything is packaged as an R package containing the analysis functions and the visualization tool as a shiny web application. The makefile, reference sequences for HPV and the HeLa example dataset are installed along the R package. The makefile allows for batch processing of multiple samples, and multithreading is enabled for bowtie2, samtools, blat, and most of the R logic.

Viral capture data are usually small. For example, the HeLa sample dataset is 13Mb per paired FASTQ file, and executes in 12 minutes using 4 threads on a standard laptop computer (Intel i7 CPU, 16GB of RAM).

The pipeline code can be obtained from https://github.com/maximewack/viroCapt. A demo instance showing the HeLa results can be found here: https://shiny.maximewack.com/viroCapt_MIE

The validation on samples with HPV, HBV, and AAV2 integrations yielded concordant, well-ranked results (Table 1), showing good retrieval performances of the algorithm and assorted filters.

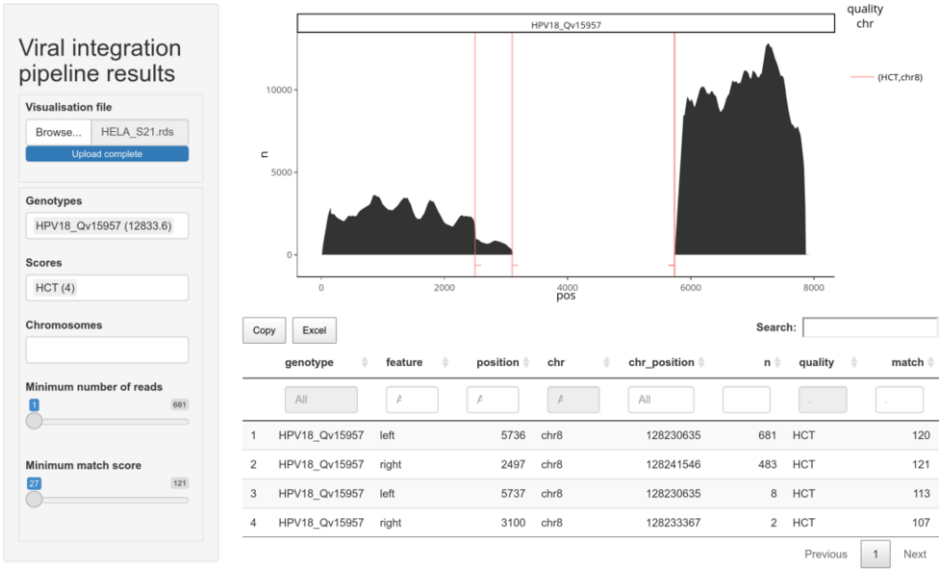


Figure 2. Screenshot of the interactive analysis tool displaying the HeLa sample results.

Table 1. Comparison of viral integration identified in viroCapt and the references

		viroCapt		**references**		
virus	**sample**	virus	human (chr:position)	virus	human	**rank**
HBV	1	1671 - 2844	19:30291247 - 30291291	1671 - 2849	30291247 - 30291287	1 - 2
		1827 - 2846	5:1295275 - 1295209	1830 - 2843	1295275 - 1295207	23 - 24
		1794 -	12:79363501 -	1789 -	79363502 -	25
		192 - 2224	6:49526654 - 49411818	189 - 2226	49526654 - 49411817	26 - 726
	2	1738 - 1824	14:90044537 - 90132716	1736 - 1828	90044536 - 90132717	1 - 3
		1496 - 587	5:1295211 - 1026640	1496 - 587	1295211 - 1026641	4 - 7
		2985 - 1864	19:30300745 - 30301945	2981 - 1853	30300749 - 30301935	14 - 15
AAV	1	4390 - 4632	5:1295307 - 1295291	4390 - 4597	1295308 - 1295291	1 - 3
	2	4571 - 4270	3:172224024 - 172224026	4270 - 4571	17222402 7 - 172224026	1 - 2
	3	*no integration*		*no integration*		
	4	*no integration*		*no integration*		
	5	4597 - 4386	3:172302190 - 172224151	4388 - 4597	172224150 - 172302191	1 - 2
HPV	HeLa	2497 - 5736	8:128241546 - 128230635	2497 - 5736	128241494 - 128230632	1 - 2
		3100 -	8:128233367 -	3088 -	128233367 -	4

4. Discussion

We compared the results obtained using our analytic pipeline to those obtained by published results on HeLa cells [11] and liver cancer samples with HPV or AAV2 integrations, [12] and most results are confirmed in the top results.

While the carcinogenic mechanisms specific to each viral insertion are still largely unknown, high throughput approaches allow the genotyping and identification of genome insertions. We already used this pipeline in clinical research to elaborate on those mechanisms [14,15].

This is not the first pipeline to implement a viral insertion finding logic, some implementing a similar strategy [16,17], but this is the first to include such a dedicated analysis tool.

The addition of an interactive web tool for the exploration of results and assisting in interpretability is a valuable asset for researchers and clinicians alike to make use of these data in an efficient manner.

References

[1] Jackson SE, Chester JD. Personalised cancer medicine. International Journal of Cancer. 2015 Jul;137(2):262-6.
[2] Parkin DM. The global health burden of infection-associated cancers in the year 2002. International Journal of Cancer. 2006 Jun;118(12):3030-44.
[3] Levrero M, Zucman-Rossi J. Mechanisms of HBV-induced hepatocellular carcinoma. Journal of Hepatology. 2016 Apr;64(1 Suppl):S84-S101.
[4] Munger K, Baldwin A, Edwards KM, Hayakawa H, Nguyen CL, Owens M, et al. Mechanisms of human papillomavirus-induced oncogenesis. Journal of Virology. 2004 Nov;78(21):11451-60.
[5] Duncavage EJ, Magrini V, Becker N, Armstrong JR, Demeter RT, Wylie T, et al. Hybrid capture and next-generation sequencing identify viral integration sites from formalin-fixed, paraffin-embedded tissue. The Journal of molecular diagnostics: JMD. 2011 May;13(3):325-33.
[6] Holmes A, Lameiras S, Jeannot E, Marie Y, Castera L, Sastre-Garau X, et al. Mechanistic signatures of HPV insertions in cervical carcinomas. NPJ Genomic Medicine. 2016;1:16004. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5685317/.
[7] Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods. 2012 Mar;9(4):357-9.
[8] Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009 Aug;25(16):2078-9.
[9] Kent WJ. BLAT–the BLAST-like alignment tool. Genome Res. 2002 Apr;12(4):656-64.
[10] Chen X, Kost J, Li D. Comprehensive comparative analysis of methods and software for identifying viral integrations. Briefings in Bioinformatics. 2019 Nov;20(6):2088-97. Available from: https://doi.org/10.1093/bib/bby070.
[11] Adey A, Burton JN, Kitzman JO, Hiatt JB, Lewis AP, Martin BK, et al. The haplotype-resolved genome and epigenome of the aneuploid HeLa cancer cell line. Nature. 2013 Aug;500(7461):207-11. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3740412/.
[12] Peneau C, Imbeaud S, La Bella T, Hirsch TZ, Caruso S, Calderaro J, et al. Hepatitis B virus integrations promote local and distant oncogenic driver alterations in hepatocellular carcinoma. Gut. 2021 Feb.
[13] La Bella T, Imbeaud S, Peneau C, Mami I, Datta S, Bayard Q, et al. Adeno-associated virus in the liver: natural history and consequences in tumour development. Gut. 2020 Apr;69(4):737-47.
[14] Morel A, Neuzillet C, Wack M, Lameiras S, Vacher S, Deloger M, et al. Mechanistic Signatures of Human Papillomavirus Insertions in Anal Squamous Cell Carcinomas. Cancers. 2019 Nov;11(12).[15]
[15] Pere H, Vernet R, Pernot S, Pavie J, Robillard N, Puech J, et al. Episomal HPV16 responsible for aggressive and deadly metastatic anal squamous cell carcinoma evidenced in peripheral blood. Scientific Reports. 2021 Feb;11(1):4633.
[16] Wang Q, Jia P, Zhao Z. VirusFinder: Software for Efficient and Accurate Detection of Viruses and Their Integration Sites in Host Genomes through Next Generation Sequencing Data. PLOS ONE. 2013 May;8(5):e64465. Publisher: Public Library of Science. Available from: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0064465.
[17] Chen Y, Yao H, Thompson EJ, Tannir NM, Weinstein JN, Su X. VirusSeq: software to identify viruses and their integration sites using next-generation sequencing of human cancer tissue. Bioinformatics. 2013 Jan;29(2):266-7. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/

Challenges of Trustable AI and Added-Value on Health
B. Séroussi et al. (Eds.)

doi:10.3233/SHTI220603

Designing an Optimal Expansion Method to Improve the Recall of a Genomic Variant Curation-Support Service

Anaïs MOTTAZ[a,b,1], Emilie PASCHE[a,b], Pierre-André MICHEL[a,b], Luc MOTTIN[a,b], Douglas TEODORO[a,b,c] and Patrick RUCH[a,b]
[a] *HES-SO / HEG Geneva, Information Sciences, Geneva, Switzerland*
[b] *SIB Swiss Institute of Bioinformatics, Geneva, Switzerland*
[c] *Department of Radiology and Medical Informatics, University of Geneva, Geneva, Switzerland*

Abstract. The importance of genomic data for health is rapidly growing but accessing and gathering information about variants from different sources is hindered by highly heterogeneous representations of variants, as outlined by clinical associations (AMP/ASCO/CAP) in their recommendations. To enable a smooth and effective retrieval of variant-containing documents from different resources, we developed a tool (https://goldorak.hesge.ch/synvar/) that generates for any given SNP – including variant not present in existing databases – its corresponding description at the genome, transcript and protein levels. It provides variant descriptions in the HGVS format as well as in many non-standard formats found in the literature along with database identifiers. We present the SynVar service and evaluate its impact on the recall of a genomic variant curation-support service. Using SynVar to search variants in the literature enables to increase the recall by +133.8% without a strong impact on precision (i.e. 93%).

Keywords. genomic variant, biomedical literature, precision medicine

1. Introduction

In the search for accelerating medical discoveries and improving health, genomic data is gaining increasing importance. Cohorts of human genomes are getting larger and cancer sequencing has become routine. Such data facilitate the study of the underlying causes of diseases, help refining care planning and treatment efficacy. But accessing and gathering genomic data and variants information from different sources are highly challenging.

Genetic variants correspond to a change from a template sequence and can be described using various reference sequences and at different levels (genome, transcript, protein). A given variant can thus correspond to several descriptions, even when using nomenclature standards as described by the HGVS society [1]. The combinatorial nature of the different description levels (many-to-many relationships), due to gene overlapping, isoforms and genetic code redundancy, hinder a linear mapping between

[1] Corresponding Author, Anaïs Mottaz, Campus Battelle, Bâtiment B, Rte de la Tambourine 17, 1227 Carouge, Switzerland; E-mail: anais.mottaz@hesge.ch.

them. While many databases of polymorphisms and variants exist, such as ClinVar, COSMIC or dbSNP, using those resources as terminologies is fairly challenging since they do not identify variants at the same level. Besides, although guidelines from clinical associations (AMP/ASCO/CAP) recommend unambiguous variants naming for interpretation and reporting [2], resources such as the biomedical literature tend to use heterogenous and non-standard ways of representing variants like missing reference sequence leading to ambiguous position, or fancy syntaxes.

To enable a smooth and effective retrieval of documents containing variants, we developed a service to expand variant queries, SynVar. This service provides for a given SNP its corresponding description at the genome, transcript and protein levels, in the HGVS format as well as in many non-standard descriptions found in the literature (e.g. BRAFV600E). The SynVar service is being used for query expansion by Variomes (https://candy.hesge.ch/Variomes/), a high recall search engine to support the curation of genomic variants [3]. It is also used by CINECA, which develops a federated infrastructure for genetic data sharing, implementing GA4GH standards, and their Beacon models to query cohorts for genetic variants over many data nodes (https://www.cineca-project.eu). We present, here, the architecture of SynVar, along with an evaluation of its impact on the recall of variant searches in the literature and the frequency distribution of the different expansion patterns encountered in the literature.

2. Methods

2.1. SynVar architecture

The main steps of the SynVar service are presented below. Depending on the description level of the queried variant, the order and the steps may slightly differ.

Query processing. Variants can be queried at the level of the protein, transcript or genome, dbSNP or COSMIC identifier. Variant is extracted from the query using regular expressions. The variant does not need to be in a standard HGVS format. The reference gene, chromosome or sequence is also extracted using regular expressions from the *ref* parameter, if given, or the *variant* parameter. If it is provided within the *variant* parameter, no specific format or order is required (e.g. "Val600Glu, BRAF"). Gene name is validated through the neXtProt API [4]. The description level of the variant is optional and guessed from its format if not provided with the query.

Variant validation. For protein and transcript variants, reference sequence identifiers corresponding to the provided gene are retrieved using the neXtProt API, including isoforms if *iso* parameter is set to true. Protein or transcript fasta sequences are retrieved from neXtProt and NCBI E-utilities [5] respectively. Reference amino acid or base of the variant is then checked against the reference sequences at the variant position. For genomic variants, validation is done against assemblies GRCh37 and GRCh38 using *variantValidator* [6]. If a gene is provided instead of a chromosome, the validation is done on the corresponding chromosome only if the position lies within the given gene.

Translation/Backtranslation. Valid protein and transcript variants in HGVS format, as generated by the validation or conversion steps, are, respectively, backtranslated using *Mutalyzer Back Translator* and translated using *runMutalyzerLight* tools [7]. Several variants may be generated by backtranslation due to amino acid code redundancy.

Mapping/Conversion. Transcript variants are mapped to GRCh37 and GRCh38 assemblies and genomic variants are converted to transcript variants using *variantValidator* or, when not available, *Mutalyzer numberConversion* service.

Variant identifiers. DbSNP identifiers are retrieved based on the chromosome reference sequence and position using NCBI e-Utilities, after the mapping to the genomic build. If a dbSNP identifier is provided as a query, the different genomic variants corresponding to the identifier are collected using NCBI e-Utilities. For COSMIC, the mapping between transcript variants and COSMIC identifiers are retrieved from the downloaded *COSMIC Mutation data* [8] and used to provide COSMIC identifiers in the output and to enable query expansion from a COSMIC identifier.

Syntactic variations. For each level of variant description, a set of syntactic variations is provided. It represents the most common variant description formats as encountered in the literature, based on previously described patterns [9] and on a preliminary evaluation of variant recognition in literature.

Output. Results are returned as a list of genomic variants, along with their corresponding transcript and protein variants in HGVS and non-standard formats. The output can be in XML or JSON Beacon format.

2.2. Evaluation

To evaluate SynVar, we performed two experiments. Both experiments are based on a set of 766 variants in BRCA1 and BRCA2. This set was built using BRCAExchange [10] and corresponds to all missense SNPs for BRCA1 and BRCA2 from LOVD [11].

First, we evaluated the effect of variant expansion generated by SynVar on the recall of a set of variant searches. Literature in PubMed Central was searched for the 766 variants using the Variomes APIs [3]. Two searches were performed for each variant: the first used only the term mentioned in the list (e.g. M18T) while the second expanded the query with all variant descriptions suggested by SynVar (e.g. 53T>C).

Second, we investigated how the different description levels and syntactic synonyms suggested by SynVar were represented in the literature. To this extent, a set of 61 expansion patterns (Table 1) was defined to represent the SynVar synonyms: 18 patterns at the genome level, 18 patterns at the transcript level and 25 patterns at the protein level. The Variomes APIs were used to retrieve all variant occurrences in PubMed Central for the 766 variants. Each occurrence was then mapped to one of the 61 patterns.

Table 1. Example of expansion patterns generated for representing SynVar variant synonyms

Pattern	Level	Example of variant
\d+[ACGT]\s*>\s*[ACGT]	Genome	43124044A > G
\d+[ACGT]/[ACGT]	Transcript	53T/C
p\.[A-Z][a-z]{2}\d+[A-Z][a-z]{2}	Protein	p.Met18Thr

3. Results

3.1. Availability

The SynVar service is available as a SaaS via an Open API (https://goldorak.hesge.ch/synvar/). The GUI is mainly for demonstration and debugging purposes. Output is available in different formats, including Beacon format.

3.2. *Impact of SynVar on variant search*

Results of the impact of SynVar on the recall for variant search are presented in Table 2. Searching variants in literature using variant expansion resulted in an improvement of the recall by 133.8%. Indeed, while a query with a single term for representing the variant returned on average 3 documents, using all SynVar descriptions enabled retrieving on average 7 documents. It also significantly reduced the number of queries without any results: without SynVar, 255 queries returned no documents, while using SynVar enabled finding documents for 118 of these 255 queries. Among documents retrieved using SynVar, 93% were relevant as estimated by a manual analysis of a subset of 27 documents retrieved from five random queries.

Table 2. Results of the comparison of Variomes with and without SynVar

	Without SynVar	**With SynVar**
Mean number of documents retrieved per query	3	7
Total number of documents retrieved for all queries	2304	5387
Number of queries with no results	255	137

3.3. *Evaluation of the SynVar variants' expansion patterns distribution*

Out of the 766 queries, 137 queries resulted in no document. In addition, for 6 queries, our system failed to map the variant occurrence to one of the expansion patterns. We thus present the results using the remaining 623 queries. 7037 variant occurrences were identified and mapped to expansion patterns. On average, a variant was represented in the literature under 3.3 different patterns (min: 1; max: 8). 60.6% of variant occurrences corresponded to a variant at the protein level, 28.2% at the transcript level and 11.2% at the genome level (Figure 1). Among the protein patterns proposed by SynVar, three patterns dominated and represented respectively 60.9%, 28.5% and 10.1% of all protein patterns. One pattern was largely represented in the transcript patterns with 94.2% of the occurrences. At the genome level, it is the dbSNP identifier which was mostly impactful with 95.1% of the occurrences.

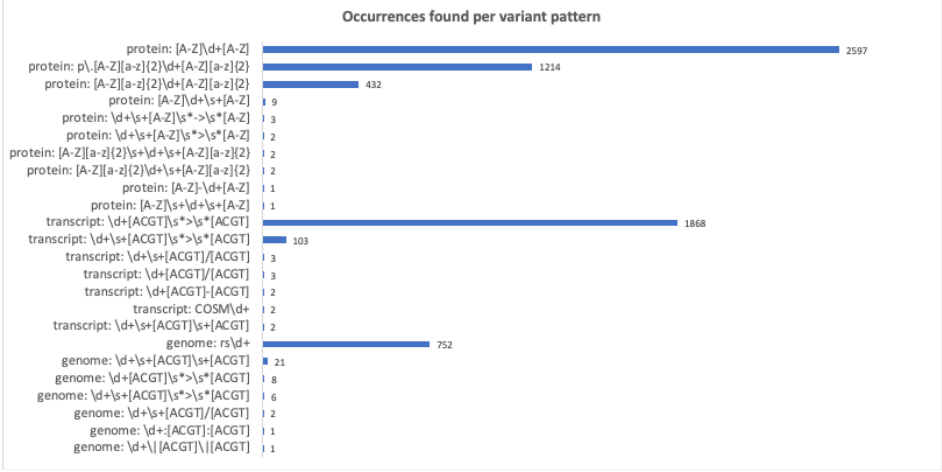


Figure 1. Number of occurrences found in the literature for each variant pattern returning matches

4. Discussion

We propose a service to facilitate the retrieval of variant-containing documents from heterogeneous resources. Its main advantage compared to existing tools, such as tmVar [12,13], is that it processes variants independently of a database, resulting in a much broader recall [3]. Indeed, using a database limits the coverage due to its specific purpose (polymorphism *vs* somatic variants database) and decreases the specificity when position-only based (like dbSNP). The impact of SynVar on literature search recall is important, as it more than doubles the number of retrieved documents without strongly altering precision (i.e. 93%). Searching for the classical pattern would allow to catch only 50% of the occurrences. While six patterns represent more than 95% of occurrences, even a few matches may be of importance for rare variants, which accounts for the vast majority of variants. In a future work, additional patterns as well as non-SNP variants will be considered along with the pre-indexing of documents to improve efficiency, using methods such as Named Entity Recognition [14] and requiring variant normalization.

Acknowledgements. This service has been supported by CINECA - UE H2020 Grant #825775 and the Canadian Institute of Health Research (CIHR) Grant #404896.

References

[1] den Dunnen JT. Describing Sequence Variants Using HGVS Nomenclature. Methods Mol Biol. 2017;1492:243-251.
[2] Li MM, Datto M, Duncavage EJ, Kulkarni S, Lindeman NI, Roy S *et al*. Standards and Guidelines for the Interpretation and Reporting of Sequence Variants in Cancer: A Joint Consensus Recommendation of the Association for Molecular Pathology, American Society of Clinical Oncology, and College of American Pathologists. J Mol Diagn. 2017 Jan;19(1):4-23.
[3] Pasche E, Mottaz A, Caucheteur D, Michel PA, Gobeill J and Ruch P. Variomes: a high recall search engine to support the curation of genomic variants. Bioinformatics.. 2022;btac146
[4] Zahn-Zabal M, Michel PA, Gateau A, Nikitin F, Schaeffer M, Audot E *et al*. The neXtProt knowledgebase in 2020: data, tools and usability improvements. Nucleic Acids Res. 2020 Jan 8;48(D1):D328-D334.
[5] Sayers E. The E-utilities In-Depth: Parameters, Syntax and More. 2009 May 29 [Updated 2021 Apr 15]. In: Entrez Programming Utilities Help [Internet]. Bethesda (MD): National Center for Biotechnology Information (US); 2010-.
[6] Freeman PJ, Hart RK, Gretton LJ, Brookes AJ, Dalgleish R. VariantValidator: Accurate validation, mapping, and formatting of sequence variation descriptions. Hum Mutat. 2018 Jan;39(1):61-68.
[7] Lefter M, Vis JK, Vermaat M, den Dunnen JT, Taschner PEM, Laros JFJ. Next Generation HGVS Nomenclature Checker. Bioinformatics. 2021 Feb 4;37(18):2811–7.
[8] Tate JG, Bamford S, Jubb HC, Sondka Z, Beare DM, Bindal N *et al*. COSMIC: the Catalogue Of Somatic Mutations In Cancer. Nucleic Acids Res. 2019 Jan 8;47(D1):D941-D947.
[9] Yip YL, Lachenal N, Pillet V, Veuthey AL. Retrieving mutation-specific information for human proteins in UniProt/Swiss-Prot Knowledgebase. J Bioinform Comput Biol. 2007 Dec;5(6):1215-31.
[10] Cline MS, Liao RG, Parsons MT, Paten B, Alquaddoomi F, Antoniou A *et al*. BRCA Challenge: BRCA Exchange as a global resource for variants in BRCA1 and BRCA2. PLoS Genet. 2018 Dec 26;14(12):e1007752.
[11] Fokkema IF, Taschner PE, Schaafsma GC, Celli J, Laros JF, den Dunnen JT. LOVD v.2.0: the next generation in gene variant databases. Hum Mutat. 2011 May;32(5):557-63.
[12] Lee K, Wei CH, Lu Z. Recent advances of automated methods for searching and extracting genomic variant information from biomedical literature. Brief Bioinform. 2021 May 20;22(3):bbaa142.
[13] Wei CH, Phan L, Feltz J, Maiti R, Hefferon T, Lu Z. tmVar 2.0: integrating genomic variant information from literature with dbSNP and ClinVar for precision medicine. Bioinformatics. 2018 Jan 1;34(1):80-87.
[14] Copara J, Naderi N, Knafou J, Ruch P, Teodoro D. Named entity recognition in chemical patents using ensemble of contextual language models. In Proceedings of the CLEF 2020 conference. 2020 Sep 22-25.

Challenges of Trustable AI and Added-Value on Health
B. Séroussi et al. (Eds.)

doi:10.3233/SHTI220604

Enriching UMLS-Based Phenotyping of Rare Diseases Using Deep-Learning: Evaluation on Jeune Syndrome

Carole FAVIEZ[a,b,1], Marc VINCENT[c], Nicolas GARCELON[a,b,c], Caroline MICHOT[d,e], Genevieve BAUJAT[d,e], Valerie CORMIER-DAIRE[d,e], Sophie SAUNIER[f], Xiaoyi CHEN[a,b,c], and Anita BURGUN[a,b,g]

[a] *Centre de Recherche des Cordeliers, Sorbonne Université, INSERM, Université de Paris, Paris, France*

[b] *Inria Paris, France*

[c] *Université de Paris, Imagine Institute, Data Science Platform, INSERM UMR 1163, Paris, France*

[d] *Reference Centre for Constitutional Bone Diseases, laboratory of Osteochondrodysplasia, INSERM UMR 1163, Imagine Institute, Université de Paris, Paris, France*

[e] *Hôpital Necker-Enfants Malades, Service de génétique, AP-HP, Paris, France*

[f] *Laboratory of Renal Hereditary Diseases, INSERM UMR 1163, Imagine Institute, Université de Paris, Paris, France*

[g] *Hôpital Necker-Enfants Malades, Département d'informatique médicale, AP-HP, Paris, France*

Abstract. The wide adoption of Electronic Health Records (EHR) in hospitals provides unique opportunities for high throughput phenotyping of patients. The phenotype extraction from narrative reports can be performed by using either dictionary-based or data-driven methods. We developed a hybrid pipeline using deep learning to enrich the UMLS Metathesaurus for automatic detection of phenotypes from EHRs. The pipeline was evaluated on a French database of patients with a rare disease characterized by skeletal abnormalities, Jeune syndrome. The results showed a 2.5-fold improvement regarding the number of detected skeletal abnormalities compared to the baseline extraction using the standard release of UMLS. Our method can help enrich the coverage of the UMLS and improve phenotyping, especially for languages other than English.

Keywords. Named entity recognition, electronic health records, deep phenotyping, rare disease

1. Introduction

The increasing digitization of health-related data provides the unique opportunity to bring new insights into disease knowledge and patient care. For rare diseases, electronic health records (EHR) provide a precious source for patient high throughput

[1] Corresponding Author, Carole Faviez, Centre de Recherche des Cordeliers, INSERM, 15 rue de l'école de médecine, F-75006, Paris, France; E-mail: carole.faviez@inserm.fr.

phenotyping that can help reduce the risks of mis- and delayed diagnosis [1]. However, the amount of unstructured data in EHRs requires that named entity recognition (NER) be developed to automatically extract medical entities from text. NER approaches include solutions leveraging thesauri to search for mentions of the terms in the documents as well as machine or deep learning. QuickUMLS [2] and cTAKES [3] are thesaurus-based solutions that have been evaluated on clinical notes in English. These systems provide normalized terms, e.g., based on the Unified Medical Language System (UMLS) CUIs [4] which enables the easy reuse and comparison of the retrieved phenotypes and the conversion to other thesauri or languages. These methods generally reach better precision than recall [2], due to the presence of specific synonyms or variations of the terms in text. Moreover, the coverage of used thesauri can vary across languages, especially for rare phenotypes. To address these issues, non-dictionary-based methods leveraging deep learning have been used to detect medical entities, such as BiLSTM-CRF [5] and PhenoTagger [6], which have been evaluated on medical articles. These developments have been part of a wider move toward deep learning in healthcare, due to the large availability of biomedical data [7]. Regarding medical entity extraction, these methods do not provide the same normalization as thesaurus-based ones and their performance on clinical text must be analyzed. In this article, we present a hybrid method to extract phenotypes from EHR documents combining a dictionary-based method [8] with deep learning. The pipeline was evaluated in a French rare disease center, with focus on Jeune syndrome (Jeune asphyxiating thoracic dystrophy), a rare genetic disorder characterized by skeletal abnormalities such as small, narrow thorax, short ribs, shortened bones of the arms and legs, unusually shaped pelvis, and extra fingers and/or toes[2].

2. Methods

2.1. Patient selection and extraction of phenotypes

The Necker Children's Hospital is a French reference center for rare and undiagnosed diseases that hosts the Imagine Institute, a research center specialized in genetic diseases. Its clinical data warehouse (Dr. Warehouse [8]) contains EHRs of more than 800,000 patients. Dr. Warehouse enables the automatic extraction of medical entities from EHRs based on the UMLS [4]. In addition, a research database is used to store structured research information from rare disease patients of Imagine Institute, including patient diagnoses. We selected all patients diagnosed with Jeune syndrome from the Imagine research database and limited ourselves to the patients followed up at Necker hospital, i.e., with EHRs in Dr. Warehouse.

UMLS extraction was performed by the high throughput phenotyping module of Dr. Warehouse [8]. The deep learning extraction was performed by applying a biGRU-CRF [9] extension to the method described in [10]. This model is well suited for NER from noisy natural language text of EHRs due to its ability to remember information across different ranges of contextual dependencies [11]. Each document was tokenized and fed as a string of fasttext word embeddings to the NER model. We reused an internal annotated dataset of 625 clinical reports in French from various departments at Necker hospital for model selection and training. These reports were manually

[2] https://rarediseases.info.nih.gov/diseases/3049/jeune-syndrome

annotated with spans corresponding to phenotypic entities (i.e., observable physical or mental characteristics of the individuals, including both pathological and normal characteristics). The annotation provided 6928 phenotypes stemming from $2.5x10^5$ tokens.

2.2. UMLS and biGRU-CRF extraction and evaluation

The set of Jeune syndrome EHRs was randomly split into two datasets: training set (two thirds of the patients) and test set (the remaining third). Phenotypes were extracted from the training set using the UMLS extraction and the biGRU-CRF method. Phenotypes detected exclusively by the biGRU-CRF method were manually reviewed to verify whether they were skeletal abnormalities associated with Jeune syndrome and mapped to UMLS concepts. The list of skeletal abnormalities and their mappings to the UMLS were validated by three bone disease experts from Necker/Imagine (VCD, GB, CM). The enriched version of the UMLS integrating these new synonyms in French will be referred to as UMLS+ in the following. The last step consisted in extracting phenotypes from the test set with the dictionary-based method using standard UMLS and UMLS+ and comparing the results. The pipeline overview is given in Figure 1.

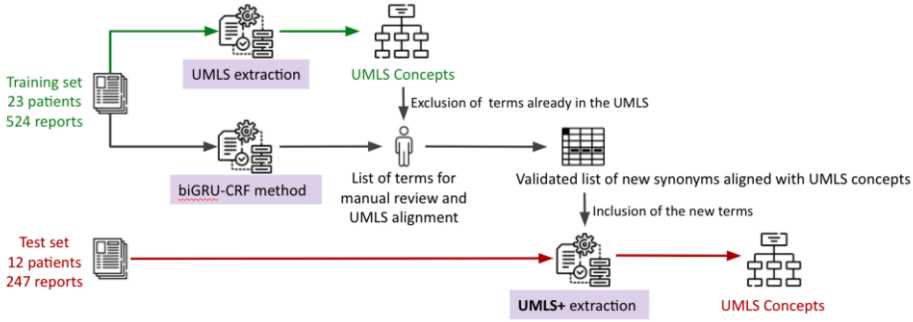


Figure 1. Overview of the pipeline.

3. Results

3.1. Training set and UMLS enrichment

Starting from the research database, we identified 142 patients with Jeune syndrome. Among them, 35 patients were followed up at Necker, with an average of 22 documents per patient in Dr Warehouse (median 10). 23 patients out of 35 were randomly selected for the training set. These patients were associated with a total of 524 documents on which UMLS extraction and biGRU-CRF techniques were applied. The standard UMLS extraction identified 592 distinct terms (3387 occurrences) in the training set. Compared with this method, the biGRU-CRF method was able to detect 1036 additional distinct terms (1922 occurrences). The review of these additional French terms by clinical experts led to the identification of 119 skeletal abnormalities related to Jeune syndrome, among which some key phenotypes such as *thoracic dystrophy*, *narrow thorax*, and *short ribs*. These 119 skeletal abnormalities were

mapped to 76 UMLS CUIs. With the addition of lexical variants (e.g., plural forms), 152 different words or expressions were added to the UMLS+. This list is available upon request to the corresponding author. This process allowed us to add French translations of English terms (e.g., *thorax etroit* is equivalent to *narrow thorax* (C0426790)) but also to identify new synonyms that were absent from the Metathesaurus in English and in French (e.g., *duplication of the fifth finger* can be mapped to *post-axial polydactyly* (C0220697)).

3.2. Comparison of UMLS and UMLS+ based extraction on the test set

Phenotypes were extracted from our test set of 12 patients and 247 documents, using the UMLS and the UMLS+. The standard UMLS extraction provided 354 distinct terms (1168 occurrences, 328 distinct CUI), including 34 skeletal abnormalities associated with Jeune syndrome (148 occurrences, 33 CUI). With UMLS+, we detected 67 skeletal abnormalities, which is to say that after enrichment, the system was able to detect 33 additional skeletal abnormalities (222 occurrences, 25 CUI). In other terms, we obtained twice as many phenotypes and a 2.5-fold increase of phenotypic information. Of note some terms are of major interest for Jeune syndrome diagnosis, like *etroitesse thoracique* (*narrow thorax*), with 21 occurrences in the test set. The most frequent phenotypes related to skeletal abnormalities are displayed in Table 1.

Table 1. Top 10 skeletal abnormalities detected in the test set with the UMLS+.

UMLS CUI	Found text (fr)	English translation	Frequency	Origin
C1406921	Dystrophie thoracique	Thoracic dystrophy	82	Added term
C0036439	Scoliose	Scoliosis	41	UMLS
C0302142	Deformation	Deformity	33	UMLS
C0426790	Etroitesse thoracique	Narrow thorax	21	Added term
C0410528	Maladie osseuse constitutionnelle	Constitutional bone disease	18	Added term
C0426817	Cotes courtes	Short ribs	17	Added term
C0022821	Gibbosite	Gibbosity	16	Added term
C0022821	Cyphose	Kyphosis	13	UMLS
C0426790	Thorax etroit	Narrow thorax	12	Added term
C1439256	Deformation thoracique	Thorax deformity	11	Added term

4. Discussion

Even though UMLS is a rich source for extracting and normalizing phenotypes from EHRs, we demonstrate in this study the need for enriching the UMLS with terms that are used in clinical documents, especially in languages other than English. We showed that deep learning techniques such as biGRU-CRF can be used as a complementary method to improve phenotyping especially in the context of rare disease. As for Jeune Syndrome, the enriched UMLS (UMLS+) enabled a 2.5-fold increase in phenotyping. Although we were able to map all the extracted phenotypes to existing CUIs, some key phenotype terms in French were totally absent from the Metathesaurus. Moreover, some refined terms were missing even in English, e.g., *duplication of the fifth finger*. These results are similar to those obtained by Vasilakes et al. [12], who compared a dietary supplement knowledge base (iDISK) with the UMLS and found that although 99% of the iDISK concepts were present in the UMLS, only 30% of the

terms/synonyms were included. Our proposed enrichment process can be easily reproduced for other rare diseases. The limitation is that a manual review of new extracted phenotypes and normalization to UMLS concepts can be time-consuming for large corpora. To address this issue, it would be interesting to develop a module for the automatic enrichment of UMLS terms. Such a system could rely on term embeddings to group terms detected from text with UMLS terms that convey similar meanings [13], as proposed by Sarker [14]. Regarding the main contributions to the field, the enrichment process that we propose provides more accurate and comprehensive phenotyping and can improve the performances of EHR-based screening of patients. This is crucial for rare diseases, as underdiagnosis and delayed diagnosis are frequent, leading to a high social and psychological burden. Future work will consist in generalizing the method to other diseases.

Acknowledgement This work was supported by the French National Research Agency (ANR) under the C'IL-LICO project (ANR-17-RHUS-0002), approved by the French ethics and scientific committee for health research (CESREES) (#2201437), and as part of the "Investissements d'avenir" (ANR-19-P3IA-0001) (PRAIRIE 3IA Institute).

References

[1] Faviez C, Chen X, Garcelon N, Neuraz A, Knebelmann B, Salomon R, et al. Diagnosis support systems for rare diseases: a scoping review. Orphanet Journal of Rare Diseases. 2020 Apr 16;15(1):94.
[2] Soldaini L, Goharian N. QuickUMLS: a fast, unsupervised approach for medical concept extraction. 2016;4.
[3] Savova GK, Masanz JJ, Ogren PV, Zheng J, Sohn S, Kipper-Schuler KC, et al. Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. J Am Med Inform Assoc. 2010 Oct;17(5):507–13.
[4] Bodenreider O. The Unified Medical Language System (UMLS): integrating biomedical terminology. Nucleic Acids Res. 2004 Jan 1;32(Database issue):D267-270.
[5] Habibi M, Weber L, Neves M, Wiegandt DL, Leser U. Deep learning with word embeddings improves biomedical named entity recognition. Bioinformatics. 2017 Jul 15;33(14):i37–48.
[6] Luo L, Yan S, Lai P-T, Veltri D, Oler A, Xirasagar S, et al. PhenoTagger: A Hybrid Method for Phenotype Concept Recognition using Human Phenotype Ontology. Bioinformatics. 2021 Jan 20;btab019.
[7] Miotto R, Wang F, Wang S, Jiang X, Dudley JT. Deep learning for healthcare: review, opportunities and challenges. Brief Bioinform. 2018 Nov 27;19(6):1236–46.
[8] Garcelon N, Neuraz A, Salomon R, Bahi-Buisson N, Amiel J, Picard C, et al. Next generation phenotyping using narrative reports in a rare disease clinical data warehouse. Orphanet Journal of Rare Diseases. 2018 May 31;13(1):85.
[9] Chung J, Gulcehre C, Cho K, Bengio Y. Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling. arXiv:14123555 [cs] [Internet]. 2014 Dec 11 [cited 2022 Jan 14]; Available from: http://arxiv.org/abs/1412.3555
[10] Vincent M, Douillet M, Lerner I, Neuraz A, Burgun A, Garcelon N. Using deep learning to improve phenotyping from clinical reports. Stud Health Technol Inform. 2021 In press;
[11] Jagannatha AN, Yu H. Bidirectional RNN for Medical Event Detection in Electronic Health Records. Proc Conf. 2016 Jun;2016:473–82.
[12] Vasilakes J, Bompelli A, Bishop JR, Adam TJ, Bodenreider O, Zhang R. Assessing the enrichment of dietary supplement coverage in the Unified Medical Language System. Journal of the American Medical Informatics Association. 2020 Oct 1;27(10):1547–55.
[13] Yuan Z, Zhao Z, Sun H, Li J, Wang F, Yu S. CODER: Knowledge-infused cross-lingual medical term embedding for term normalization. J Biomed Inform. 2022 Jan 4;126:103983.
[14] Sarker A. LexExp: a system for automatically expanding concept lexicons for noisy biomedical texts. Bioinformatics. 2021 Aug 25;37(16):2499–501.

Challenges of Trustable AI and Added-Value on Health
B. Séroussi et al. (Eds.)

doi:10.3233/SHTI220605

Classification of Oncology Treatment Responses from French Radiology Reports with Supervised Machine Learning

Jean-Philippe GOLDMAN[a,d,1], Luc MOTTIN[b,c], Jamil ZAGHIR[a,d], Daniel KESZTHELYI[a,d], Belinda LOKAJ[a], Hugues TURBÉ[a,d], Julien GOBEIL[b], Patrick RUCH[b,c], Julien EHRSAM[a] and Christian LOVIS[a,d]
[a] *Division of Medical Information Sciences, Geneva University Hospital, Switzerland*
[b] *HES-SO/HEG Genève, Information Sciences, Geneva, Switzerland*
[c] *SIB Text Mining, Swiss Institute of Bioinformatics, Switzerland*
[d] *Department of Radiology and Medical Informatics, University of Geneva, Switzerland*

Abstract. The present study shows first attempts to automatically classify oncology treatment responses on the basis of the textual conclusion sections of radiology reports according to the RECIST classification. After a robust and extended manual annotation of 543 conclusion sections (5-to-50-word long), and after the training of several machine learning techniques (from traditional machine learning to deep learning), the best results show an accuracy score of 0.90 for a two-class classification (non-progressive vs. progressive disease) and of 0.82 for a four-class classification (complete response, partial response, stable disease, progressive disease) both with Logistic Regression approach. Some innovative solutions are further suggested to improve these scores in the future.

Keywords. oncology, treatment response, RECIST, automatic classification, supervised machine learning, natural language processing

1. Introduction

The main objective of the SPO project (Swiss Personalized Oncology) funded by the SPHN (Swiss Personalized Health Network) is to develop nation-wide infrastructure for personalized oncology and to maximize benefit for the patients. A part of this consists in developing a continuous, high quality clinical and molecular data collection throughout Switzerland as well as providing tools for the best personalized care for patients.

Standardizing radiology reports for the evaluation of response to treatment across institutes implies to establish a pipeline mining unstructured texts in the electronic health record (EHR) and to extract knowledge such as diagnosis, tumor type, primary tumor site, response to treatment, and morphologic description (texture, dimension, location).

This paper describes the efforts to automatically classify the response to treatment from the conclusion section of radiology reports in French from the Geneva University Hospitals (HUG) and the Lausanne University Hospital (CHUV). The two main goals are to annotate existing EHRs with the RECIST classification (Response Evaluation

[1] Corresponding Author, Jean-Philippe GOLDMAN; E-mail: Jean-Philippe.Goldman@unige.ch

Criteria in Solid Tumours) [1] in order to build a standardized research database, and to provide a reliable decision support tool.

The results of the assessment follow the RECIST 1.1 classification with these 4 categories: Complete Response (CR), Partial Response (PR), Stable Disease (SD), Progressive Disease (PD). We also aim at classifying the same data in two classes P/NP: P = progressive tumor (PD), and NP = non-progressive tumor (CR, PR or SD).

2. Material and Methods

2.1. Data Extraction

The data used for the training of the automatic classifier come from various places and stages of the SPO project. In an early exploratory step of the project, both oncology teams of CHUV and HUG were asked to manually extract approximately 120 radiology reports, including 6 major tumor cancer types (Breast, Abdominal, Lung, Prostate, Melanoma, Blastoma), preferably with a balanced distribution of cases across the 4 RECIST classes and across the different types of cancer Several radiology modalities for assessment were included (MRI, CT-SCAN, PET-CT). HUG extraction led to 122 radiology reports (HUG122), while CHUV extracted 118 reports (CHUV118). The conclusion of these reports have 31 words in average (min=5, max=54). Here are 3 short examples: 1) Majoration de la condensation lobaire inférieure gauche, 2) Stabilité de la maladie tumorale, 3) Pas de récidive tumorale locale ou à distance

At a second stage of the project, a cohort of patients with a BRAF gene mutation was selected for other longitudinal studies, leading to a group of 108 HUG unique patients representing 303 radiology reports (BRAF303).

The reports were de-identified before processing in accordance with the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule's De-Identification Standard. The study was approved by SwissEthics (2020-00347).

2.2. Data Annotation

The three sub-corpuses (HUG122, CHUV118, BRAF303) were annotated the same way by 2 to 3 expert annotators from the respective institutions (HUG for HUG122 and BRAF303, and CHUV for CHUC118) on the basis of the conclusion section of the reports only and using the guidelines defined by oncologists involved in the SPO project, and according to the RECIST 1.1 criteria. Additionally, the annotators were asked to add two information with a yes/no label:

- *dissociated response*, for reports with at least two different RECIST classes.
- *low confidence*, when doubt or uncertainty was expressed in the report.

After the annotation of HUG122 corpus by 3 experts, the Cohen's kappa inter-annotator agreement (IAA) was calculated for each pairs of annotators (0.80, 0.79, 0.93) as well as the global IAA (0.84). The computed IAAs of the two other corpus (CHUV118 and BRAF303) were 0.88 and 0.83 respectively.

Eventually, all of the parallel annotations were reviewed by a single expert annotator from HUG, who solved annotation disagreements, yielding to a gold-standard reference corpus of 543 annotated report conclusions with RECIST classification, as described in Table 1, as well as *dissociated response*, and *low confidence* information. The two-class

annotation P (progressive) vs. NP (non-progressive) was derived from the RECIST annotation with the label P for PD conclusions, and the label NP for the others.

Table 1. Number of conclusions per RECIST class and per sub-corpus

RECIST	HUG122	CHUV118	BRAF303	Total
Source	HUG	CHUV	HUG	
1 = CR	22	12	93	127
2 = PR	21	30	33	84
3 = SD	36	20	52	108
4 = PD	40	56	110	206
5 = Unknown	3	0	15	18
Total	122	118	303	543

The 18 conclusions labelled as 5 (equivalent to *unknown* in the RECIST classification) are left out for this study. Most of these are conclusions of radiology reports that are not depicting the evolution of an oncology disease, such as the initial assessment. Among the remaining 525 conclusions (hereafter named as *complete dataset*), there are 19 conclusions annotated as *dissociated response*, while the conclusions flagged as low confidence represent 21 items, with 3 conclusions in common (i.e. annotated both as *dissociated response* and *low confidence*). All in all, there are 488 *non-ambiguous* conclusions. This latter set of 488 conclusions is named *filtered dataset*.

2.3. Automatic Classification

The complete dataset was preprocessed with the usual steps of NLP pipelines:

- *character normalization,* (but diacritics were kept) ;
- *removal of stop-words,* (the words implying *negation* were excluded from the stop-word list. In other words, these *negation* words were kept in the resulting preprocessed dataset in order to keep the valence of a phrase to have a clear distinction between *progression vs. no progression*) ;
- *lemmatization,* (using the well-known NLP framework SpaCy [2] together with *fr_core_news_md*, a language model trained on news in French) ;
- use of *CountVectorizer* (a procedure that converts a collection of text documents to a matrix of token counts. Some preliminaries attempts showed the best results with counts combining words, 2-grams and 3-grams of words).

Four techniques from traditional machine learning (ML) domains [3] and two deep learning (DL) techniques were selected and trained with the data.

1. SVM-RBF: A support vector machines (SVM) technique is a supervised learning method mainly used for classification, outliers detection and regression. It has the advantage of being effective in high dimensional spaces. Also it uses a subset of training points in the decision function (called support vectors), for memory efficiency purpose. It allows various kernel function for the decision function like for example the Radial Basis Function (RBF).
2. SVM-Lin: The Linear SVM is a similar technique using a linear decision function. The Stochastic Gradient Descent (SGD) approach is using a convex loss function, which was proven to have good results with large-scale and sparse machine learning problems found in NLP.

3. NB: The Naive Bayes method is also a supervised learning algorithm based on the Bayes' theorem with the naive assumption of conditional independence between every pair of features given the value of the class variable. [4]
4. LR: the Logistic Regression is a multiclass classifier using the one-vs-rest scheme, also known as logit or Maximum Entropy regression. [5]

The two deep learning techniques used in this study were:

1. Feed-forward neural network (FNN): the FNN includes an embedding layer (in our configuration, with embedding bag of size 128), a mean over vectors and a linear layer (of size 128) outputting an array of size *num_class*. *num_class* represents the number of classes (2 for the P-NP task, 4 for the RECIST task in our case). The FNN was trained with a CrossEntropy Loss function with a learning rate of 10^{-2}. To avoid overfitting, an early stopping strategy (ESS) was used to stop the training as soon as the validation accuracy reached a ceiling.
2. Convolutional network (CNN): the CNN also includes an embedding layer, several layers of Conv-2D (4 layers in our configuration with kernel sizes of (2,128) to (5,128)), a linear layer (of size 512, and an output array of size *num_class*) and a Softmax layer. It was also trained with a CrossEntropy Loss function with the ESS, with a batch size of 16, and a learning rate of 10^{-5}.

Both NN trainings used 4-fold cross-validation (i.e. the test is 25% of the dataset). During the training, 20% of the training set is used as validation set.

3. Results

The data were used to train the 6 classifiers for the targeted categories (RECIST and P-NP) with and without filtering out the conclusions labeled as *dissociated response* or *low confidence*. The *complete dataset* (i.e. non-filtered) included 525 conclusions while the *filtered dataset* included 488 conclusions. The following table shows the accuracy obtained for all the configurations.

Table 2. Accuracy of the 6 learning techniques for complete and filtered dataset and for the two classification tasks (RECIST, PNP)

	complete dataset (n=525)		filtered dataset (n=488)	
	RECIST	PNP	RECIST	PNP
SVM-RBF	0.74	0.87	0.77	0.88
NB	0.75	0.83	0.76	0.84
SVM-Lin	0.79	0.85	0.80	0.87
LR	**0.82**	**0.88**	**0.82**	**0.90**
FNN	0.77	0.85	0.81	0.89
CNN	0.78	**0.88**	0.80	0.89

Table 2. shows similar results across all 6 techniques, ranging from 0.74 for the RECIST task with the complete dataset and NB approach, to 0.82 for LR. Comparing with inter-annotator agreement as mentioned in section 2.2, LR is also the best choice for the 2-class PNP task (with an accuracy of 0.88). As for the *filtered dataset* (n=488), the removal of the 18 ambiguous conclusions yielded slightly better results as expected with an accuracy of 0.82 (respectively 0.90) for the RECIST task (resp. for the PNP task). Figure 1. shows the confusion matrix for the LR approach with the well-classified

conclusions on the diagonal and the erroneous classifications (18 items among 102) evenly distribute. Some surprising cases show a mix up between radically opposite labels (Complete Response vs. Progressive Disease). Additional qualitative introspection should be done to diagnose these wrong classifications.

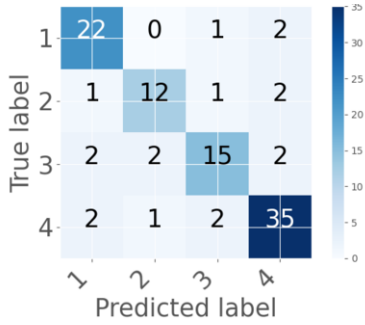


Figure 1. Average confusion matrix for the validation set (20% of the complete dataset) with the LR approach for the RECIST classification tasks (1=CR, 2=PR, 3=SD, 4=PD)

4. Discussion

The two main results of this study are the followings. First, the resulting accuracy is always comparable to the human inter-annotator agreement. Second, across the variety of ML approaches tests in this study, Logistic Regression has the best results for all configurations (*complete* vs *filtered dataset*, and RECIST vs PNP). Also, one can notice that the DL approaches (FNN and CNN) while performing worse than the LR approach, have comparable results with the other ML approaches. The surprisingly equivalent performances observed with DL techniques can possibly be explained by the small size of the corpus. Using the results, several other studies could be designed. The confusion matrix suggests to implement some other target task such as regression or ordinal classification [6] (instead of classification), to take in account the hypothetic proximity of consecutive classes (i.e. CR (1) conclusions are more alike PR (2) than PD (4)).

Additionally, using a larger dataset, the accuracy variation across institutes (CHUV vs. HUG) could be tested. Finally, one should try to reproduce the classification task at the sentence level instead of whole conclusions as the latter can describe different evolutions on various tumor sites and get better accuracy scores.

Acknowledgements: This work was funded by the Swiss Personalized Health Network – SPHN.

References

[1] Eisenhauer EA, et al. New response evaluation criteria in solid tumours: revised RECIST guideline (version 1.1). Eur J Cancer. 2009 Jan;45(2):228-47.
[2] Honnibal, M., & Montani, I. (2017). spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing.
[3] Buitinck et al. (2013). API design for machine learning software: experiences from the scikit-learn project
[4] H. Zhang (2004). The optimality of Naive Bayes. Proc. FLAIRS.
[5] C.M. Bishop (2006) Pattern Recognition and Machine Learning, Chapter 4.3.4
[6] Frank, E., & Hall, M.A. (2001). A Simple Approach to Ordinal Classification. ECML.

Challenges of Trustable AI and Added-Value on Health
B. Séroussi et al. (Eds.)

doi:10.3233/SHTI220606

Clustering Nursing Sentences - Comparing Three Sentence Embedding Methods

Hans MOEN[a,1], Henry SUHONEN[b], Sanna SALANTERÄ[b,c], Tapio SALAKOSKI[d], and Laura-Maria PELTONEN[b]
[a]*Department of Computer Science, Aalto University, Espoo, Finland*
[b]*Department of Nursing Science, University of Turku, Turku, Finland*
[c]*Turku University Hospital, Turku, Finland*
[d]*Department of Mathematics and Statistics, University of Turku, Turku, Finland*

Abstract. In health sciences, high-quality text embeddings may augment qualitative data analysis of large amounts of text by enabling, e.g., searching and clustering of health information. This study aimed to evaluate three different sentence-level embedding methods in clustering sentences in nursing narratives from individual patients' hospital care episodes. Two of these embeddings are generated from language models based on the BERT framework, and the third on the Sent2Vec method. These embedding methods were used to cluster sentences from 20 patient care episodes and the results were manually evaluated. Findings suggest that the best clusters were produced by the embeddings from a BERT model fine-tuned for the proxy task of predicting subject headings for nursing text.

Keywords. Text clustering, natural language processing, sentence embeddings, nursing documentation, electronic health records

1. Introduction

Vectorized representations (embeddings) of text that captures meaning in a semantic space are important for many tasks related to natural language processing (NLP). This includes searching, clustering, summarization, and classification. The interest in textual embeddings has rapidly been growing since the introduction of the Word2Vec method [1]. The current direction focuses on contextualized embeddings with pre-trained language models like ELMo [2], and more recently transformer-based models like BERT [3]. Studies show that BERT-based language models without fine-tuning for a downstream task perform quite poorly, often underperforming compared to averaging the word embeddings from traditional (global) word embedding methods [4]. Thus, fine-tuning of BERT models on relevant tasks and domains seems important for generating embeddings that capture the semantics of the targeted domain.

In health sciences, text clustering may augment qualitative data analysis of large amounts of health-related text to support both clinical work as well as research. Our aim is to group sentences from nursing notes from individual patients' hospital stays (care episodes) into clusters that each focus on one and the same topic, or possibly a coherent set of topics for sentences that cover more than one topic.

[1]Corresponding author, Hans Moen, Tietotekniikan laitos, P.O. Box. 15400, FI-00076 AALTO, Finland; E-mail: hans.moen@aalto.fi.

We explore performance of three sentence-level embedding methods. We limited the experiment to using the same clustering approach for all three. Given that our focus is clinical text, which differs from the type of text used when pre-training the used BERT model [5], we hypothesize that the fine-tuning of the model on a proxy classification task with nursing text will yield better embeddings. We also test the Sent2Vec [6] method, which has shown strong performance in unsupervised word and sentence embedding tasks.

2. Methods

2.1. Data

The clinical data set used in this study consists of nursing documents from electronic health records (2005-2020) of almost 94,000 cardiac patients from a Finnish hospital district. Ethical (hospital ethics committee 17.2.2009 §67; UTU ethics committee 9/2020) and administrative approvals were obtained (2/2009; J14/20). In this hospital district, nurses structure their text according to the Finnish Care Classification standard (FinCC) [7], which is a taxonomy (with >600 headings) of nursing diagnoses, interventions and outcomes. We used a subset of about 1 million nursing documents (3.4 million sentences) for model training. Another subset of 20 care episodes was used in the manual evaluation (135 documents, 1032 paragraphs, 2301 sentences).

2.2. Automatic Clustering

We used k-means clustering [8] which is a centroid-based algorithm because it indicated best performance in an initial pilot evaluation. For implementation we used the PyClustering library [9]. Other clustering algorithms tested in the pilot study were OPTICS [10], DBSCAN [11] and Agglomerative Hierarchical clustering (see, e.g., [12]). The Euclidean distance gave better or, at worst, similar results as Cosine similarity based on the pilot test. Since an optimal number of clusters was unknown, we used an automatic approach for this. Both OPTICS and DBSCAN are designed to solve this problem. However, none of these two outperformed k-means clustering together with a technique for determining optimal cluster number. We calculated the Silhouette coefficient for the different number of clusters (k) and picked the one that had the highest coefficient (see e.g. [13]). We used the implementation in Scikit-learn [14]. Another technique considered for determining the optimal cluster number was the elbow technique [15, 16].

2.3. Sentence Embeddings

We explored three embedding methods for generating sentence embeddings:

- `BERT-BASIC` is the BERT model pre-trained for Finnish text on news, online discussions, and internet crawls [5]. When inputting a sentence, the embedding is extracted from the representation of the '[CLS]' token in its last layer.
- `BERT-FINE_TUNED` is the Finnish BERT model further fine-tuned on the nursing text dataset for a proxy task focusing on classifying nursing sentences based on subject headings in the FinCC standard (consists of > 600 headings)

(see [17] for more information about this task). Machine learning libraries used are Huggingface [19], PyTorch [20] and Keras/Tensorflow [21].

- SENT2VEC is an embedding model trained using the Sent2Vec method [6]. It learns static word n-gram embeddings, using an approach similar to C-BOW in Word2vec [1] where the context window equals sentence length. These are combined into sentence embeddings. We trained this on the mentioned nursing text dataset using default parameters except dim=200 and loss=hs. We did not incorporate the FinCC headings here.

2.4. Manual Evaluation

A common way to evaluate clustering results is to manually create gold clusters, and then use a cluster similarity score like the Rand index [18] to compare the generated clusters against. We found manually forming gold clusters very difficult due to the complexity and size of the data. Instead, it was easier for domain experts to evaluate clustering results retrospectively. Thus, we formulated an evaluation scheme where evaluators were instructed to score each cluster according to two criteria - topic coherency (Evaluation A), and uniqueness of topic(s) (Evaluation B). See Table 1. For evaluation B, cases with less than 4 clusters in a care episode, evaluation scores of 2 or 3 were changed to 4 retrospectively to not favor very large and few clusters. A sample of 20 care episodes was used in the manual evaluation (see Data section). Evaluators were two specialists in nursing. Interrater agreement was calculated with Cohen's Kappa.

Table 1. Evaluation scheme used for scoring the clustering results by the different methods

Evaluation A Topic coherency of each cluster		**Evaluation B** Uniqueness of the topic(s) to each cluster	
Class	**Description**	**Class**	**Description**
1-ideal	All sentences cover the same topic(s).	1-ideal	Topic(s) are unique to this cluster.
2-semi-optimal	One topic found here is not in all sentences.	2-semi-optimal	Same topic(s) also occurs in one other cluster.
3-poor	Two of the topics here are not found in all sentences.	3-poor	Same topic(s) occur in two other clusters.
4-very bad	Three or more of the topics are not found in all sentences.	4-very bad	Same topic(s) occur in three or more clusters.
5-unable to assess	-	5-unable to assess	-

3. Results

Evaluation scores are shown in Table 2 (Evaluation A) and Table 3 (Evaluation B). We report the scores on sentence level to compensate for differences in cluster sizes. The BERT-FINE_TUNED method outperformed both BERT-BASIC and SENT2VEC with a larger number of clusters (more sentences) rated class 1 and 2 (ideal and semi-optimal): For the topic cluster coherency evaluation criteria (A), 77.53-78.45% of clusters formed by BERT-FINE_TUNED belong to classes 1 or 2. For BERT-BASIC this number is 59.24-62.53%, while for SENT2VEC it is 34.72-35.59%. When it comes to evaluating the uniqueness of the topic(s) to each cluster, criteria (B), 16.82-26.90% of clusters formed by BERT-FINE_TUNED belong to classes 1 or 2, while this number is 4.19-7.48% for BERT-BASIC, and 11.12-16.04% for SENT2VEC.

Table 2. Scores from manual evaluation A - topic coherency of each cluster.

	SENT2VEC				BERT-BASIC				BERT-FINE_TUNED			
Evaluator:	**1**		**2**		**1**		**2**		**1**		**2**	
Class	*n*	%	*n*	%	*n*	%	*n*	%	*n*	%	*n*	%
1	689	30	643	28	1036	45	1060	46	1037	45	1092	47
2	130	6	156	7	403	17	303	13	747	32	713	31
3	135	6	61	3	242	10	132	6	231	10	178	8
4	1345	58	1397	61	612	27	749	33	279	12	259	11
5	2	.1	44	2	8	.4	57	2	7	.3	59	3

Table 3. Scores from manual evaluation B - uniqueness of the topic(s) to each cluster.

	SENT2VEC				BERT-BASIC				BERT-FINE_TUNED			
Evaluator:	**1**		**2**		**1**		**2**		**1**		**2**	
Class	*n*	%	*n*	%	*n*	%	*n*	%	*n*	%	*n*	%
1	18	.8	173	7	17	.7	64	3	75	3	323	14
2	238	10	196	8	156	7	49	2	381	17	298	13
3	176	8	223	10	223	10	26	1	537	23	74	3
4	1830	79	1546	67	1872	81	2075	90	1278	55	1546	67
5	39	2	163	7	33	1	87	4	30	1	60	3

The interrater agreement varied between methods and evaluation criteria. For SENT2VEC the overall agreement was *moderate* (A: 0.55, n=489, $p<0.05$; B: 0.43, n=489, $p<0.05$). For BERT-BASIC the agreement was only *fair* to *moderate* (A: 0.47, n=761, $p<0.05$; B: 0.30, n=761, $p<0.05$). Also, for BERT-FINE_TUNED the agreement was *fair* to *moderate* (A: 0.43, n=533, $p<0.05$; B: 0.21, n=553, $p<0.05$).

4. Discussion

As BERT-FINE_TUNED gives the most suited sentence embeddings for this task, this highlights the importance of domain and task specificity in the fine-tuning of these models. For specialized domains in healthcare there are usually very few task-specific labeled datasets available. This study shows that, by formulating a proxy classification task on the data and labels that are available, we can still fine-tune generic language models to better represent the semantics of such specialized text. The evaluation showed that it is an easier task to generate coherent clusters compared to generating clusters with no topical overlap. This mirrors well the holistic nature of nursing documentation, where different aspects of care are interconnected. The interrater evaluation scores confirm the difficulty of determining what constitutes a coherent cluster and how to discern between intercluster overlap. Our scores could also indicate that the interpretation of the evaluation scheme differs somewhat between the evaluators. Further research is needed to build a theoretical framework for clustering and evaluating nursing text as well as for validating the findings of this study with a larger sample.

Study limitations include limited evaluation sample and modest evaluation scheme. The focus of this study was not to find the optimal clustering algorithm or distance metric. Still, better results can likely be achieved through a more thorough evaluation of different clustering algorithms, embedding methods, distance metrics and techniques for determining optimal cluster counts. However, this requires a different evaluation approach than what we have used. Additionally, future research should consider explainability aspects and assess confidence of methods used.

5. Conclusions

Contextual sentence embeddings generated by a BERT model fine-tuned on a proxy classification task shows promising results when used for clustering nursing text from cardiac patients' narratives. The findings can be used to develop tools to augment health science researchers in qualitative analysis of large data sets. As future work we plan to test these embeddings for the purpose of extractive summarization of nursing text.

Acknowledgements: This work was supported by the Academy of Finland (grants 315376, 336033, 315896), Business Finland (grant 884/31/2018), and EU H2020 (grant 101016775). We thank Jari Björne for helping with fine-tuning the BERT model.

References

[1] Mikolov T, Chen K, Corrado G, Dean J. Efficient estimation of word representations in vector space. arXiv:1301.3781 [cs.CL]. 2013 Jan 16.
[2] Peters ME, Neumann M, Iyyer M, Gardner M, Clark C, Lee K, Zettlemoyer L. Deep contextualized word representations. arXiv:1802.05365 [cs.CL]. 2018 Feb 15.
[3] Devlin J, Chang MW, Lee K, Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In: Proceedings of NAACL 2019. 2019 June; Minneapolis, Minnesota. 4171-86.
[4] Jiang T, Huang S, Zhang Z, Wang D, Zhuang F, Wei F, et al. PromptBERT: Improving BERT Sentence Embeddings with Prompts. arXiv:2201.04337 [cs.CL]. 2022 Jan 12.
[5] Virtanen A, Kanerva J, Ilo R, Luoma J, Luotolahti J, Salakoski T, et al. Multilingual is not enough: BERT for Finnish. arXiv:1912.07076 [cs.CL]. 2019 Dec 15.
[6] Pagliardini M, Gupta P, Jaggi M. Unsupervised Learning of Sentence Embeddings Using Compositional n-Gram Features. In: Proceedings of NAACL 2018. 2018, June. New Orleans: Louisiana. 528-40.
[7] Hoffrén P, Leivonen K, Miettinen M. Nursing standardized documentation in Kuopio University Hospital. Stud Health Technol Inform 2009;146:776-7.
[8] Lloyd S. Least squares quantization in PCM. IEEE Trans Inf Theory 1982;28(2):129-37.
[9] Novikov AV. PyClustering: Data mining library. J Open Source Softw 2019;4(36),1230.
[10] Ankerst M, Breunig MM, Kriegel HP, Sander J. OPTICS: Ordering points to identify the clustering structure. ACM Sigmod record 1999;28(2),49-60.
[11] Ester M, Kriegel HP, Sander J, Xu X. A density-based algorithm for discovering clusters in large spatial databases with noise. In: KDD-96 Proceedings. 1996, Aug. 226-231.
[12] Jain AK, Murty MN, Flynn PJ. Data clustering: a review. ACM Comput Surv 1999 Sept;31(3),264-323.
[13] Frahling G, Sohler C. A fast k-means implementation using coresets. Int J Comput Geom Appl 2006 Jan;18(06),605-25.
[14] Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine learning in Python. J Mach Learn Res 2011 Oct;12,2825-30.
[15] Thorndike RL. Who belongs in the family? Psychometrika 1953 Dec;18(4),267-76.
[16] Syakur MA, Khotimah BK, Rochman EMS, Satoto BD. Integration k-means clustering method and elbow method for identification of the best customer profile cluster. IOP Conference Series: Materials Science and Engineering 2018 Apr;336(1),012017.
[17] Moen H, Hakala K, Peltonen LM, Suhonen H, Ginter F, Salakoski T, et al. Supporting the use of standardized nursing terminologies with automatic subject heading prediction: a comparison of sentence-level text classification methods. JAMIA 2020;27(1),81-8.
[18] Hubert L, Arabie P. Comparing partitions. J Classif 1985 Dec;2(1)193-218.
[19] Wolf T, Debut L, Sanh V, Chaumond J, Delangue C, Moi A, et al. Huggingface's transformers: State-of-the-art natural language processing. arXiv:1910.03771 [cs.CL]. 2019 Oct 9.
[20] Paszke A, Gross S, Massa F, Lerer A, Bradbury J, Chanan G, et al. PyTorch: An imperative style, high-performance deep learning library. In: Advances in neural information processing systems. 2019 Dec 8-14. Vancouver: Canada. 8026-37.
[21] Abadi M, Agarwal A, Barham P, Brevdo E, Chen Z, Citro C, et al. Tensorflow: Large-scale machine learning on heterogeneous distributed systems. arXiv:1603.04467 [cs.DC]. 2016 Mar 14.

Challenges of Trustable AI and Added-Value on Health
B. Séroussi et al. (Eds.)

doi:10.3233/SHTI220607

Discovering Key Topics in Emergency Medical Dispatch from Free Text Dispatcher Observations

Pablo FERRI[a,1], Carlos SÁEZ[a], Antonio FÉLIX-DE CASTRO[b], Purificación SÁNCHEZ-CUESTA[b] and Juan M GARCÍA-GÓMEZ[a]
[a] *Biomedical Data Science Laboratory (BDSLab), Instituto de Aplicaciones de las Tecnologías de la Información y de las Comunicaciones Avanzadas (ITACA), Universitat Politècnica de València (UPV), Valencia, Spain*
[b] *Conselleria de Sanitat Universal i Salut Pública, Generalitat Valenciana (GVA), Valencia, Spain*

Abstract. The objective of this work was to discover key topics latent in free text dispatcher observations registered during emergency medical calls. We used a total of 1374931 independent retrospective cases from the Valencian emergency medical dispatch service in Spain, from 2014 to 2019. Text fields were preprocessed to reduce vocabulary size and filter noise, removing accent and punctuation marks, along with uninformative and infrequent words. Key topics were inferred from the multinomial probabilities over words conditioned on each topic from a Latent Dirichlet Allocation model, trained following an online mini-batch variational approach. The optimal number of topics was set analyzing the values of a topic coherence measure, based on the normalized pointwise mutual information, across multiple validation K-folds. Our results support the presence of 15 key topics latent in free text dispatcher observations, related with: ambulance request; chest pain and heart attack; respiratory distress; head falls and blows; fever, chills, vomiting and diarrhea; heart failure; syncope; limb injuries; public service body request; thoracic and abdominal pain; stroke and blood pressure abnormalities; pill intake; diabetes; bleeding; consciousness. The discovery of these topics implies the automatic characterization of a huge volume of complex unstructured data containing relevant information linked to emergency medical call incidents. Hence, results from this work could lead to the update of structured emergency triage algorithms to directly include this latent information in the triage process, resulting in a positive impact in patient wellbeing and health services sustainability.

Keywords. Medical emergencies, emergency medical calls, emergency medical dispatch, natural language processing, topic discovery, latent dirichlet allocation.

1. Introduction

Emergency medical dispatch entails the reception and management of demands for medical assistance in an emergency medical services system [1]. It involves emergency medical calls attendance and events triage according to their priority, process generally managed by emergency medical dispatchers. These mediators tend to follow a clinical protocol focused on a small set of structured clinical variables [2].

[1] Corresponding Author, E-mail: pabferb2@upv.es.

In the Valencian Community (Spain), the triage of emergency medical call incidents (EMCI) is currently assisted by an in-house triage protocol, a clinical decision tree based on the collection of structured variables. The dispatcher raises questions to the caller until reaching a final tree node, which has a priority assigned to it, the incident priority.

However, information not covered by the decision tree is also registered during the call in an unstructured manner in free text fields. This information, complementary to that provided by the structured variables, cannot be taken into account automatically by the clinical protocols, and thus, it is left unused.

We have studied in previous works that considering these free text dispatcher observations notably improves EMCI triage. Specifically, we have developed DeepEMC[2], a deep learning model able to automatically deal with structured and unstructured information in real time, providing performance increases of 12.5%, 17.5% and 5.1% in terms of macro F1-score in life-threatening, admissible response delay and emergency system jurisdiction prediction, respect to the current in-house triage protocol of the Valencian emergency medical dispatch service [3].

In addition, prior studies have shown the potential of text mining techniques and, concretely, topic extraction methods, to infer high-level information from huge amounts of unstructured medical data [4], [5].

Given the utmost relevance of free text dispatcher observations in EMCI triage and the availability of methods to explore them from a machine learning perspective, we present in this work an unsupervised analysis of these free text fields, with the aim of 1) discovering and understanding what information dispatchers report during emergency call incidents and 2) exploring how this latent information is distributed across incidents.

2. Methods

A total of 1374931 free text dispatcher observations linked to EMCI of the Health Services Department of the Valencian Community, were compiled in retrospective from 2014 to 2019. Data use was approved by the Institutional Review Board of the GVA and any information that may disclose the identity of the patient was discarded prior to any analysis. Given the data source, our available free text fields were written in Spanish.

A set of preprocessing operations were carried out in order to reduce dimensionality to enhance posterior topic extraction processes. Dispatcher observations were converted to lowercase and then, as text fields were written in Spanish, accents marks were deleted. Punctuation marks were also discarded along with stopwords. Words not appearing at least 50 times in the corpus were dropped, resulting in a vocabulary reduction from 74914 to 4584—discarding 94% of terms— while keeping around 96% of the total word counts in the corpus. Finally, text fields were tokenized.

Data was split using a holdout [6] methodology, with proportions of 80% for training and then 20% for testing. Next, cross-validation [6] splits were conducted over the training set, taking K=4, without allowing repetition.

Topics were inferred from the multinomial probabilities over words conditioned on each topic from a Latent Dirichlet Allocation (LDA) [7] model. We preferred LDA over Latent Semantic Analysis [8] because LDA offers a generative modeling approach, and LDA over Probabilistic Latent Semantic Analysis [9] (PLSA) because the number of parameters estimated in PLSA grows linearly with the number of training documents and generalization to new documents is easier with LDA. LDA is a hierarchical generative Bayesian model, which assumes the existence of K latent topics in a collection of text

documents. Next we present the generative process of LDA, to generate a corpus D of M documents each one with N_d words:

1. Choose $N_d \sim Poisson(\xi)$.
2. Choose $\theta_d \sim Dirichlet(\alpha)$.
3. For each of the N_d words w_{di}:
   a. Choose a topic index $z_{di} \sim Multinomial(\theta_d)$, $z_{di} \in \{1, \ldots, K\}$.
   b. Choose a word w_{di} from $p(w_{di}|z_{di}, \beta)$, a multinomial probability conditioned on the topic z_{di}.

The LDA model was trained following an online mini-batch variational inference approach [10]. The optimal number of topics was set analyzing the values of a topic coherence measure, where word context vectors were created using the normalized pointwise mutual information [11]. The distance among word context vectors was calculated with the cosine distance, obtaining the final coherence score as the arithmetic mean of all distances, following the procedures described in [12].

We tested different number of topics, specifically 5, 10, 15, 20, 25, 30 and 35. For each combination, we trained four LDA models, one per training K-fold, and calculated the aforementioned topic coherence measure in their respective validation folds. That number of topics offering the best overall performance across the validation K-folds was considered as the optimal number of topics.

Finally, we retrained the model with all the training data using the optimal configuration. For each topic, the most probable words were extracted and studied to infer topic semantics and naming it. After that, we derived the topic distribution in the training and the test corpora, to understand which were the most frequent topics in dispatcher free text fields, as well as to evaluate potential overfitting issues.

3. Results

Figure 1 shows the value of the topic coherence performance metric across the different K-folds, for each number of topics combination:

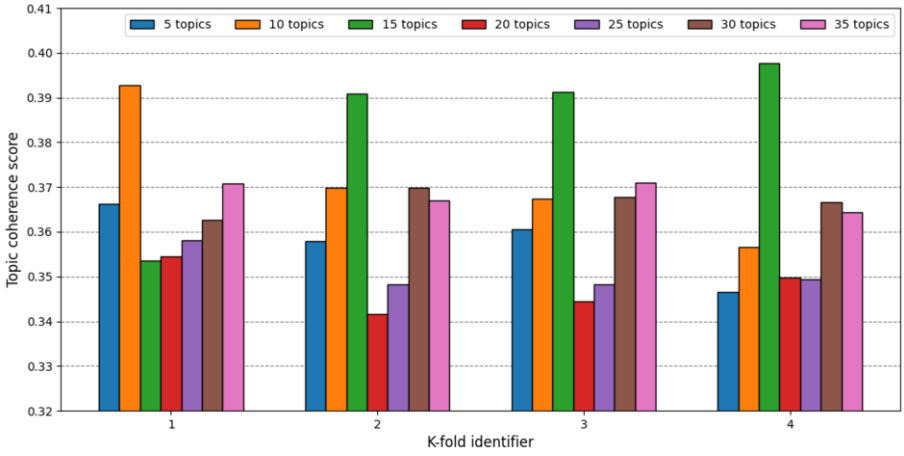


Figure 1. Number of topics selection. Topic coherence across K-folds over training set.

It can be appreciated that the optimal topic coherence value is reached at 15 topics.

Figure 2 displays the 8 words with higher associated probability respect to each topic. Each topic has been named according to the semantics its defining words represent:

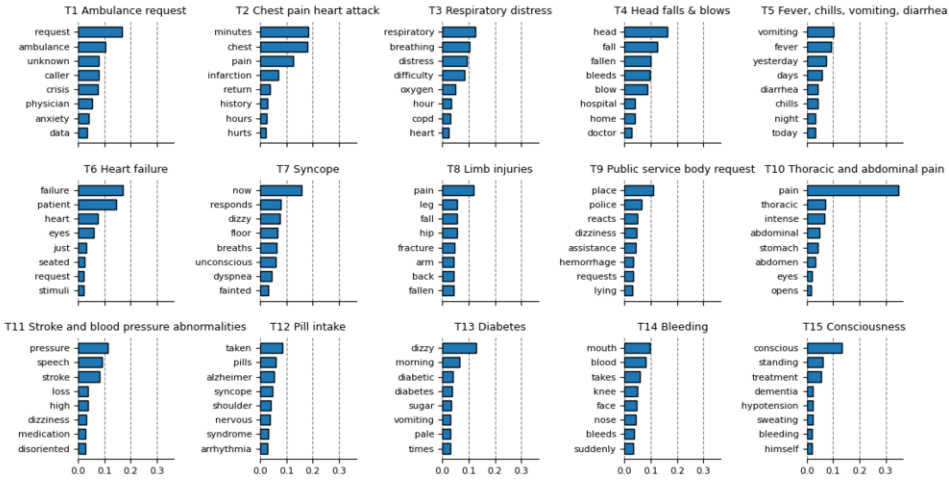


Figure 2. Topics discovered, described by their 8 words with highest probability conditioned on each topic. Word probabilities conditioned on each topic are represented in the x-axis.

It can be observed the presence of 13 clinical topics—T2, T3, T4, T5, T6, T7, T8, T10, T11, T12, T13, T14, T15—along with 2 resource dispatch topics—T1, T9. Likewise, most predominant semantics in the clinical topics are cardiovascular disorders—T2, T6, T11—and injuries T4, T8, T14.

Figure 3 presents the distribution of topics in the training and test corpora, sorted by its frequency of appearance in descending order:

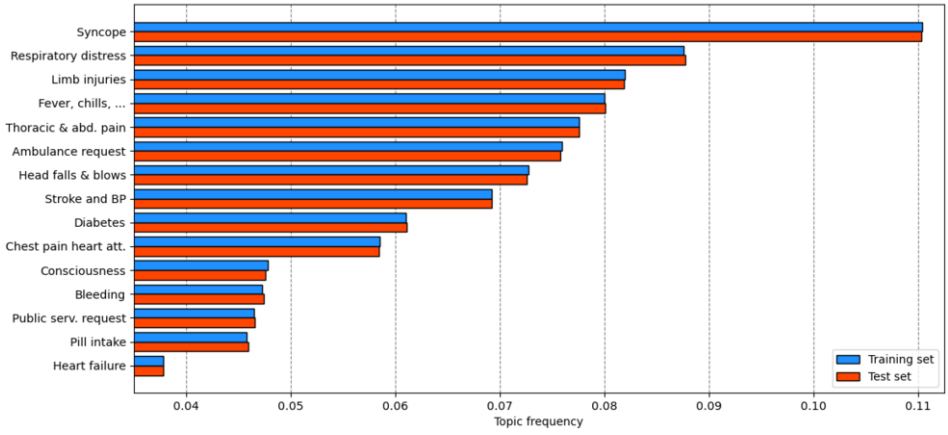


Figure 3. Topics distribution in the train and test corpus, sorted by frequency in descending order.

It can be inferred from this figure that there are not over-represented or under-represented topics. Likewise, there is a strong similarity between training and test topic distributions. Both are good signs indicating that overfitting does not seem to be present.

4. Discussion

The characterization of casuistry latent in complex unstructured data carried out in this work may lead emergency medical professionals to redefine structured decision tree algorithms in order to improve emergency medical dispatch processes.

Although the majority of topics are well-defined and delimited, some topics would require further study to evaluate the presence of topic mixtures and subtopics.

Future work will include studying relations among the topics found and potential clusters bound to the structured variables also registered during the incident. Finally, it is of interest to study why some words appear in different contexts, i.e., topics, despite having similar meanings, such as chest pain and thoracic pain.

5. Conclusions

This work has tackled the discovery of key topics in emergency medical dispatch from free text dispatcher observations. A pipeline comprising word filtering operations, number of topics selection and Latent Dirichlet Allocation model training, has been applied over 1374931 independent retrospective cases from the Valencian emergency medical dispatch service in Spain. Results support the existence of 15 latent topics, whose consideration could lead to the improvement of clinical triage protocols, deriving in turn, in a positive impact in patient wellbeing and health services sustainability.

References

[1] J. J. Clawson and K. B. Dernocoeur, *Principles of emergency medical dispatch*. Salt Lake City, Utah: Priority Press, 2003.

[2] S. J. Stratton, "Triage By Emergency Medical Dispatchers," *Prehosp. Disaster med.*, vol. 7, no. 3, pp. 263–269, Sep. 1992, doi: 10.1017/S1049023X00039601.

[3] P. Ferri *et al.*, "Deep ensemble multitask classification of emergency medical call incidents combining multimodal data improves emergency medical dispatch," *Artificial Intelligence in Medicine*, vol. 117, p. 102088, Jul. 2021, doi: 10.1016/j.artmed.2021.102088.

[4] X. Cheng, Q. Cao, and S. S. Liao, "An overview of literature on COVID-19, MERS and SARS: Using text mining and latent Dirichlet allocation," *Journal of Information Science*, p. 0165551520954674, Aug. 2020, doi: 10.1177/0165551520954674.

[5] J. Pérez, A. Pérez, A. Casillas, and K. Gojenola, "Cardiology record multi-label classification using latent Dirichlet allocation," *Computer Methods and Programs in Biomedicine*, vol. 164, pp. 111–119, Oct. 2018, doi: 10.1016/j.cmpb.2018.07.002.

[6] R. Kohavi, "A study of cross-validation and bootstrap for accuracy estimation and model selection," in *Proceedings of the 14th international joint conference on Artificial intelligence - Volume 2*, San Francisco, CA, USA, Aug. 1995, pp. 1137–1143.

[7] D. M. Blei, "Latent Dirichlet Allocation," p. 30.

[8] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman, "Indexing by latent semantic analysis," *J. Am. Soc. Inf. Sci.*, vol. 41, no. 6, pp. 391–407, Sep. 1990, doi: 10.1002/(SICI)1097-4571(199009)41:6<391::AID-ASI1>3.0.CO;2-9.

[9] T. Hofmann, "Probabilistic Latent Semantic Indexing," p. 8.

[10] M. Hoffman, F. R. Bach, and D. M. Blei, "Online Learning for Latent Dirichlet Allocation," p. 9.

[11] M. Röder, A. Both, and A. Hinneburg, "Exploring the Space of Topic Coherence Measures," in *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*, Shanghai China, Feb. 2015, pp. 399–408. doi: 10.1145/2684822.2685324.

[12] S. Syed and M. Spruit, "Full-Text or Abstract? Examining Topic Coherence Scores Using Latent Dirichlet Allocation," in *2017 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, Oct. 2017, pp. 165–174. doi: 10.1109/DSAA.2017.61.

Challenges of Trustable AI and Added-Value on Health
B. Séroussi et al. (Eds.)

doi:10.3233/SHTI220608

Topic Modeling for International Patients' Consultations Using Natural Language Processing

Sunmi LEE[a], Sookkyung SUNG[a], Seo Young KANG[a], Sujin HA[a] and Eunjoo JEON[b,1]
[a]*International Healthcare Center, Asan medical center, Seoul, Republic of Korea*
[b]*Technology Research, Samsung SDS, Seoul, Republic of Korea*

Abstract. We extracted major topic by applying natural language processing and keyword extracting using TF, TF-IDF, TextRank, Yake, KeyBERT. 1452 consultation data were collected from the website and official hospital e-mail. We found six topics categorized into "Medical opinion" related to hospital characteristics and "Non-medical service guidance". Based on this result, it is necessary to establish marketing plan and develop a digital solution for effective consultation.

Keywords. International patients, text consultations, Natural Language Processing, Topic Modeling.

1. Introduction

With the expansion of the global healthcare industry, the number of international patients entering Korea to receive treatment has increased [1]. As such, Korean hospitals were asked various questions related to medical travel thorough the text consultation. The purpose of this study was to organize and elicit meaning from these inquiries and understand the diverse needs of international patients.

2. Methods

From January 1, 2019, to May 31, 2021, 1,452 text consultation in Mongolian, Russian, Chinese, or English received from either website or official hospital e-mail. The text data were translated into English, and then went through the Natural Language Processing as tokenizing, removing stopwords, lemmatization by using Python NLTK library. Next, keyword extraction was performed in five ways (TF, TF-IDF, TextRank, Yake, KeyBERT) and the BERT model was chosen because it reflects accurate meaning with high performance [2]. Lastly, clustering was conducted by the fifty most frequent keywords using Latent Dirchlet Allocation and the final subjects were derived by six health professionals including physician, nurse, and coordinator.

[1] Corresponding Author, Eunjoo Jeon, Technology Research, Samsung SDS, Seoul, Republic of Korea; E-mail: yesredpig@gmail.com.

3. Results

Six topics were initially identified after summarizing the expert's responses (Table 1).

Table 1. Category, Topics, and Tokens Identified in Text consultation for international patients

Category	Topics	Tokens
Medical opinion	The possibility of liver transplant	liver, suffer, available, treatment, availability, transplant, possible, transplantation, surgical
	Treatment for cancer disease	cancer, chemotherapy, tumor, available, treatment, want, carcinoma, metastasis, treat
	Confirming the exact diagnosis	diagnosis, diagnose, thyroid, breast, lung, blood, disease, pain, ultrasound, biopsy, surgical, diabetes, kidney, fracture, review, bone
Non-medical service guidance	Hospital information	email, clinic, appointment, consultation, surgeon, hospital, professor, doctor
	Length of stay	long, stay, plan, therapy, test, expect, hospitalization
	Estimated cost of treatment	cost, estimate, much, examination, surgical, need, check, therapy, medicine, ultrasound, test

4. Discussion

The main reason for medical opinion category is that the hospital has achieved excellent performance in the field of liver transplantation and cancer treatment [3, 4]. In addition, non-medical inquiry about hospital information, treatment period, and estimated cost was revealed, which can be considered essential at the preliminary text consultation of medical tourism [5].

5. Conclusion

To identify international patients' inquiries to Korean hospitals, this study extracts hidden information from online consultations using NLP. By applying NLP, it is possible to reduce time consuming work and provide objective indicators to establish marketing plan and develop a digital solution (e.g., chatbot) for effective consultation in distant.

References

[1] STATISTA Inc (2021). *Medical tourism in South Korea.* Retrieved January,2022. from https://www.statista.com/study/72966/medical-tourism-in-south-korea/

[2] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

[3] Kim S-H, Hwang S, Song G-W, Jung D-H, Ahn C-S, Kim K-H, et al. Institutional standard framework and experience of living donor liver transplantation for overseas non-Korean patients at Asan Medical Center. 2021;35(2):93-9.

[4] 2022 NEWSWEEK DIGITAL LLC (2022). *World's Best Specialized Hospitals 2022*. Retrieved January 2022. from https://www.newsweek.com/worlds-best-specialized-hospitals-2022/oncology

[5] Eom T, Yu J, Han H. Medical tourism in Korea – recent phenomena, emerging markets, potential threats, and challenge factors: a review. Asia Pacific Journal of Tourism Research. 2019;24(6):563-73. doi: 10.1080/10941665.2019.1610005.

Challenges of Trustable AI and Added-Value on Health
B. Séroussi et al. (Eds.)

doi:10.3233/SHTI220609

Deep Learning-Based Brain Hemorrhage Detection in CT Reports

Gıyaseddin BAYRAK [a,1], Muhammed Şakir TOPRAK [b,c] Murat Can GANİZ [a]
Halife KODAZ [c] and Ural KOÇ [b]

[a] *Computer Engineering Department, Marmara University, Turkey*
[b] *Ministry of Health, Turkey*
[c] *Computer Engineering Department, Konya Technical University, Turkey*

Abstract. Radiology reports can potentially be used to detect critical cases that need immediate attention from physicians. We focus on detecting Brain Hemorrhage from Computed Tomography (CT) reports. We train a deep learning classifier and observe the effect of using different pre-trained word representations along with domain-specific fine-tuning. We have several contributions. Firstly, we report the results of a large-scale classification model for brain hemorrhage detection from Turkish radiology reports. Second, we show the effect of fine-tuning pre-trained language models using domain-specific data on the performance. We conclude that deep learning models can be used for detecting brain Hemorrhage with reasonable accuracy and fine-tuning language models using domain-specific data to improve classification performance.

Keywords. NLP, Deep Learning, Brain Hemorrhage, Radiology

1. Introduction and Methodology

There are several studies that develop classification models for radiology reports. For example, [1] studied Epilepsy classification using bi-LSTM on a small dataset of radiology reports from MRI. Our contributions can be summarized as follows: 1) An implementation for the critical non-traumatic hemorrhage detection from radiology reports. 2) A comparison between the baseline pre-trained FastText [2] and BERT [3] language models and task-specific fine-tuned variations of them. 3) One of the first studies to train deep learning models on a large textual dataset consist of Turkish radiology reports. [2]

The brain CT reports are labeled using the 10_{th} version of the International Classification of Diseases (ICD-10) diagnostic codes. There are two ICD-10 codes assigned to each report, the first for preliminary diagnosis, assigned before the diagnostic imaging, and the second for the final diagnosis after examining the images and radiology report. We use about 100,000 brain CT reports for patients with a preliminary diagnosis of hemorrhage indicated with I60, I61, I62. The reports whose preliminary and final di-

[1] Corresponding Author: Istanbul, Turkey; E-mail: giyaseddinalfarkh@marun.edu.tr.

[2] The use of data is approved under ethics vote number E.Kurul-E1-22-2326 by the ethical committee of the Ministry of Health. Points of view in this document are those of the authors and do not necessarily represent the official position or policies of the Ministry of Health of the Government of Turkey.

Pre-trained Embedding Layer	$\text{Precision}_{Positive}$	$\text{Recall}_{Positive}$	Macro F-1
1) $FastText_{common\ crawl}$[3]	$79.06_{\pm 0.71}$	$21.81_{\pm 0.98}$	$55.17_{\pm 0.64}$
- $FastText_{common\ crawl\ (frozen)}$	$74.82_{\pm 3.42}$	$23.98_{\pm 2.18}$	$55.98_{\pm 0.94}$
2) $FastText_{pre-trained\ on\ 190k\ unsupervised\ reports}$	$80.85_{\pm 1.09}$	$21.25_{\pm 0.42}$	$54.98_{\pm 0.32}$
- $FastText_{pre-trained\ on\ 190k\ unsupervised\ reports\ (frozen)}$	$78.58_{\pm 2.11}$	$23.03_{\pm 1.10}$	$55.92_{\pm 0.52}$
3) $Randomly\ initialized\ embeddings_{No\ pre-trained\ weights}$	$68.16_{\pm 1.56}$	$50.35_{\pm 4.21}$	$66.89_{\pm 0.95}$
4) $BERT_{Base}$[4]	$72.92_{\pm 2.50}$	$47.49_{\pm 2.60}$	$67.53_{\pm 0.41}$
5) $BERT_{Fine-tuned\ on\ trainingreports}$	$71.93_{\pm 2.82}$	$57.76_{\pm 3.88}$	$71.07_{\pm 0.55}$
6) $BERT_{Fine-tuned\ on\ 190k\ unsupervised\ reports}$	$74.39_{\pm 3.98}$	$58.66_{\pm 6.44}$	$\mathbf{72.21}_{\pm 0.97}$

Table 1. Precision, Recall and F1 scores of bi-LSTM classifier with different word representations.

agnoses codes match are labeled as positive (15697), the rest as negative (21819). Our training, validation, test split ratios are 64, 16, and 20, respectively. For pre-training and fine-tuning of language models (LM) we use 190,000 brain and thorax CT reports. We use the same deep learning classifier [4] bi-LSTM with fixed hyper-parameters with different word representations [5] to see the effect of pre-trained models and fine-tuning. Bi-LSTM is trained for 4 epochs using ADAM optimizer with a learning rate of 0.001.

2. Results and Conclusion

We show the effect of using different word representations in Table 1. As can be seen in the table, the choice of word representation, e.g. fastText or BERT and its fine-tuning has a drastic effect on the performance of the classifier. As expected, BERT contextual embeddings work better than static embeddings of fastText. Fine-tuning BERT pre-trained model with smaller but labeled task-specific data makes a difference. The difference is most visible when we fine-tune BERT with larger amounts of unsupervised domain-specific data.

We conclude that deep learning models can be used for detecting Brain Hemorrhage from radiology reports with reasonable performance ($\simeq$ 72% F1) and fine-tuning pre-trained language models such as BERT using domain-specific data to improve classification performance. In the future, we plan to develop a highly accurate and explainable real-world classification system to detect critical brain hemorrhage.

References

[1] Bayrak S, Yucel E, Takci H. Epilepsy Radiology Reports Classification Using Deep Learning Networks. COMPUTERS, MATERIALS AND CONTINUA : Tech Science Press. 2022;70(2):3589-607.

[2] Bojanowski P, Grave E, Joulin A, Mikolov T. Enriching word vectors with subword information. Transactions of the Association for Computational Linguistics. 2017;5:135-46.

[3] Devlin J, Chang MW, Lee K, Toutanova K. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:181004805. 2018.

[4] Bustos A, Pertusa A, Salinas JM, de la Iglesia-Vayá M. Padchest: A large chest x-ray image dataset with multi-label annotated reports. Medical image analysis. 2020;66:101797.

[5] Drozdov I, Forbes D, Szubert B, Hall M, Carlin C, Lowe DJ. Supervised and unsupervised language modelling in Chest X-Ray radiological reports. Plos one. 2020;15(3):e0229963.

[3]The publicly available fastText model https://dl.fbaipublicfiles.com/fasttext/vectors-crawl/cc.tr.300.bin.gz

[4]The publicly available model https://huggingface.co/dbmdz/bert-base-turkish-uncased

Challenges of Trustable AI and Added-Value on Health
B. Séroussi et al. (Eds.)

doi:10.3233/SHTI220610

Automatic Prediction of Semantic Labels for French Medical Terms

Thierry HAMON[a,1] and Natalia GRABAR[b]
[a] *Université Paris-Saclay, CNRS, LISN, F-91400, Orsay, France*
Université Sorbonne Paris Nord, F-93430, Villetaneuse, France
[b] *CNRS, Univ. Lille, UMR 8163 - STL - Savoirs Textes Langage, F-59000 Lille, France*
hamon@limsi.fr, natalia.grabar@univ-lille.fr

Abstract. We address the problem of semantic labeling of terms in two French medical corpora with the subset of the UMLS. We perform two experiments relying on the structure of words and terms, and on their context: 1) the semantic label of already identified terms is predicted; 2) the terms are detected in raw texts and their semantic label is predicted. Our results show over 0.90 F-measure.

Keywords. Semantic labeling, terminology, NLP, Machine Learning, French

1. Introduction

Semantic labeling of terms consists of assigning semantic type (e.g. disorders, procedures, medication, chemical components, anatomy, signs and symptoms) to a term which is given or identified in the documents. Recent initiatives motivated the research on semantic labeling, mainly on English medical texts for concept normalization [1,2], but also French medical texts [3]. We aim to predict semantic labels of medical terms, with or without the detection of their boundaries within texts. We rely on two corpora which are part of the CLEAR corpus [4]: the French Wikipedia (1,324 articles, 3M words) and Cochrane database (7,678 abstracts, 4.5M words). The 238,983 French terms of the UMLS [5] associated with one of the 15 semantic groups [6] are projected on the corpora. The annotations are used as reference data (respectively, 58,213 and 123,880 recognized terms).

2. Methods and Results

On the basis of the reference data, we perform two experiments. The word or term features are related to their structure (the inflectional form, prefixes and suffixes with 1 to 3 characters, presence of uppercased and lowcased characters, and presence of special characters and numbers) and to their context (inflectional forms, lemmas and POS-tags within the 5-word windows on the left and on the right. We evaluate the results with 10-

[1] Corresponding Author, Thierry Hamon, LISN, Campus universitaire bât 507, Rue du Belvedère, F - 91405 Orsay cedex, France; E-mail: hamon@limsi.fr.

This work was partly funded by the French National Agency for Research (ANR) as part of the CLEAR project (Communication, Literacy, Education, Accessibility, Readability), ANR-17-CE19-0016-01.

fold cross-validation and standard macro-measures at the level of semantic groups: Precision, Recall, and F-measure.

In the first experiment, the terms are already identified in texts. The task is to predict their semantic group. The terms are classified through the 15 semantic groups with several algorithms (Decision Trees [7], Random Forest [8], and SVM [9]). The SVM provides the best results: it outperforms Random Forest by 0.30, for instance, and gives balanced values of Precision and Recall. The average of the performance for all semantic groups with SVM is above 0.94 in both corpora. Cochrane abstracts get slightly better results than Wikipedia articles. Results indicate that it is quite easy to differentiate the 15 semantic groups among them on the basis of term structure and context. The second experiment consists in classifying each word according to 60 tags to predict the term boundaries and their semantic group with a BILOU representation. We used several algorithms (CRF [10], BiLSTM-CRF [11], Multilayer Perceptron (MLP) [12]). CRF outperforms BiLSTM-CRF by 0.30 and MLP by 0.40. The neural approaches are outperformed by CRF certainly because they may require larger datasets for training. The average of the CRF performance remains very high as well, with over 0.93 F-measure, while we observe that the size of classes is important. This experiment also permits to find out the most probable sequences of classes.

3. Conclusions and Future Work

We presented experiments on the semantic labeling of terms in French medical corpora through 15 semantic groups from the UMLS. The best results are obtained with CRF (over 0.90 F-measure) which identifies the boundaries of terms within documents, and predicts semantic groups of terms. In future work, we will enrich the reference dataset with adjectival forms of terms, use a BERT model for the semantic labeling and use these predictions for helping the automatic text simplification.

References

[1] Luo YF, Sun W, Rumshisky A. MCN: A comprehensive corpus for medical concept normalization. J Biomed Inform. 2019;92:103132.
[2] Henry S, Wang Y, Shen F, Uzuner O. The 2019 National Natural language processing (NLP) Clinical Challenges (n2c2)/Open Health NLP (OHNLP) shared task on clinical concept normalization for clinical records. J Am Med Inform Assoc. 2020;27(10):1529-37.
[3] Wajsbürt P, Sarfati A, Tannier X. Medical concept normalization in French using multilingual terminologies and contextual embeddings. J Biomed Inform. 2021;114:103684.
[4] Grabar N, Cardon R. CLEAR – Simple Corpus for Medical French. In: Workshop on Automatic Text Adaption (ATA); 2018. p. 1-11.
[5] Lindberg D, Humphreys B, McCray A. The Unified Medical Language System. Methods Inf Med. 1993;32(4):281-91.
[6] McCray AT, Burgun A, Bodenreider O. Aggregating UMLS semantic types for reducing conceptual complexity. In: Proceedings of Medinfo. vol. 10; 2001. p. 216-20.
[7] Quinlan J. C4.5 Programs for Machine Learning. San Mateo, CA: Morgan Kaufmann; 1993.
[8] Breiman L. Random Forests. Machine Learning. 2001;45(1):5-32.
[9] Cortes C, Vapnik V. Support-Vector Networks. In: Machine Learning; 1995. p. 273-97.
[10] Lafferty JD, McCallum A, Pereira FCN. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In: ICML'01. 2001. p. 282-9.
[11] Hochreiter S, Schmidhuber J. Long short-term memory. Neural computation. 1997;7(8):1735-80.
[12] Rosenblatt F. The Perceptron: a probabilistic model for information storage and organization in the brain. Psychological Review. 1958;65(6):386-408.

Challenges of Trustable AI and Added-Value on Health
B. Séroussi et al. (Eds.)

doi:10.3233/SHTI220611

Integration of Annotated Phenotype, Gene and Chemical Text Data to Advance Exposome Informatics

Christopher HAWTHORNE[a] and Guillermo H. LOPEZ-CAMPOS[a,1]
[a] *Wellcome-Wolfson Institute for Experimental Medicine, Queen's University Belfast, Belfast, UK*

Abstract. The field of phenomics has a range of biomedical informatics tools such as the Human Phenotype Ontology, providing a structured vocabulary with relationships between abnormal phenotype terms. Artificial intelligence has been widely used for entity extraction and tagging large corpora of text from PubMed and is reflected in applications such as PheneBank and PubTator. Phexpo is a tool for predicting chemical – phenotype relationships and vice-versa, although lacks the ability to decipher known relationships from unknown. Integration of these three resources can provide new meaningful relationships between phenotypes, genes and chemicals and has yet to be fully leveraged. Here we present a methodology to construct two new datasets for phenotype – gene and phenotype – chemical relationships and showcase how these datasets can be used to enhance exposome informatics.

Keywords. Phenotypes, Chemicals, Text-mining, Data Integration, Bioinformatics.

1. Introduction

Diseases and their relationship and interactions with human health is a continual research question. Diseases can routinely be broken down into their constituent phenotypes (the observable physical characteristics) in the example of COVID-19, which at the start of the pandemic was characterized by a continuous cough, fever and fatigue. Phenotypes and their individual relationships to other entities such as chemicals or genes is of great biomedical informatics interest. The Human Phenotype Ontology (HPO) provides a gold standard ontology of phenotype terms as well as associated genes derived from disease-phenotype associations. Text-mining resources such as PheneBank [1] sought to annotate the entirety of PubMed with phenotypic annotations from HPO, other resources which annotate the entirety of PubMed are PubTator [2] which provides genes and chemical annotation. Phexpo [3], is a previously published tool by the group which predicts potential chemical and phenotype relationships using their overlapping genes. Although this tool provides chemical and phenotype predictions, the results have no clear indication if relationships are known or unknown in the scientific literature and have to be validated through manual literature retrieval. The leveraging and integration of these resources has yet to be fully realized. Hence, we utilize these resources to present a new methodology to construct phenotype-gene relationships and phenotype-chemical relationships utilizing gold standard text-mining data from PheneBank and PubTator to further expand phenotype-gene and phenotype-chemical relationships to compare to

HPO's phenotype gene dataset and integrate them with phexpo to validate predictions as well as highlight novel relationships respectfully to further exposome informatics.

2. Methods

PheneBank data was preprocessed to filter out annotation scores greater than the 1st quartile based on first annotations and joined with PubTator gene or PubTator chemical datasets (2020-02-15) in R based on PubMed ID. A further quality control step was applied considering only phenotype – gene relationships sharing at least 3 PubMed IDs. An example phenotype of 'Osteoporosis' and a FDR p.value adjusted threshold of 0.05 were used for phexpo and the phenotype – chemical dataset was compared to the results.

3. Results

Phenotype – gene joining produced 12,272 phenotypes associated to genes. Compared against HPO's phenotype-gene dataset (2021-02-08), we found 1872 phenotypes that were not present and could be associated to genes based on co-occurrence (shared the same PMID). Additionally, we extracted 5,928,876 pairs of phenotype – chemical literature co-occurrences and when these were compared with the results from phexpo where we had an overlap of 611 chemicals, 1,183 chemicals significantly predicted in phexpo only and 2,092 relationships only occurring in the literature.

4. Conclusions

We have shown a new methodology for a novel data integration approach of phenotype-gene and phenotype-chemical interactions, this has facilitated the expansion of the relationships among these concepts. The phenotype-chemical relationships have been combined with the results generated using phexpo validating some of the relationships and highlighting those relationships identified by phexpo that may have not been explored yet in the literature, as phexpo has the potential to predict chemical associations for over 5,000 phenotypes, this provided an approach to validate known relationships and highlight unstudied relationships on a major scale.

References

[1] Pilehvar MT, Bernard A, Smedley D, Collier N. PheneBank: a literature-based database of phenotypes. Bioinformatics. 2021. doi:10.1093/bioinformatics/btab740

[2] Wei CH, Kao HY, Lu Z. PubTator: a web-based text mining tool for assisting biocuration. Nucleic Acids Res. 2013;41: 518–522. doi:10.1093/nar/gkt441

[3] Hawthorne C, Simpson DA, Devereux B, López-Campos G. Phexpo: a package for bidirectional enrichment analysis of phenotypes and chemicals. JAMIA Open. 2020;3: 173–177. doi:10.1093/jamiaopen/ooaa023

Challenges of Trustable AI and Added-Value on Health
B. Séroussi et al. (Eds.)

doi:10.3233/SHTI220612

Extreme Gradient Boosting Based Improved Classification of Blood-Brain-Barrier Drugs

Manuskandan SUBHA RAMAKRISHNAN [a], and Nagarajan GANAPATHY [b,1]
[a] *Department of Chemical Engineering, ACT Campus, Anna University, Chennai, India*
[b] *PLRI Institute of Medical Informatics of TU Braunschweig and Medical Hannover School*

Abstract. In this study, the analysis based on boosting approach namely linear and tree method are explored in extreme gradient boosting (XGBoost) to classify blood brain barrier drugs using clinical phenotype. The clinical phenotype features of BBB drugs are Public available SIDER dataset. The clinical features namely drug's side effect, drug's indication and the combination is fed to XGBoost. Results shows that the proposed approach is able to discriminate BBB drugs. The combination of XGBoost with tree boosting is found to be most accurate (F1=78.5%) in classifying BBB drugs. This method of tree boosting in XGBoost may be extended to access the drugs for precision medicine.

Keywords. Drug Discovery, Extreme gradient Boosting, Blood-Brain-Barrier-Drugs

1. Introduction

Machine learning methods is commonly used to investigate the drug discovery and drug re-purposing for improvised precision medicine. Blood brain barrier (BBB) permeability based drugs are highly important for neurological disorder prevention cure [1,2]. However, characterization and identification of BBB based drugs using clinical phenotypes is highly challenging [1]. Recently, extreme gradient boosting methods have been reported for improved classification for multi-set features.In this work, an extreme gradient boosting (XGBoost) approach in drug prediction of BBB permeability using clinical phenotype are explored.

2. Materials and Methods

2.1. Materials (Database)

For this study, the data-set reported in Doniger et al. [2], is considered the dataset contains 91 samples in total with 38 samples are characterised as BBB permeability true and the

[1]Corresponding Author: Nagarajan Ganapathy, PLRI Institute of Medical Informatics of TU Braunschweig and Hannover Medical School, Germany; E-mail: nagarajan.ganapathy@plri.de.

53 samples are characterized as BBB permeability false. Each samples have features namely side effect (SE) and Indicators (Ind). It uses SIDER datasets (SIDER) which have been proved the BBB permeability true or false in the clinic [1,2].

2.1.1. Method

Clinical phenotype features of BBB drugs namely SE and Ind are fed to XGboost. It contains deviation of the model and regular term to prevent over-fitting as objective function with optimization [3]. The result of the sample is predicted by set of regression tree with weights of the leaves and number of leaves. Regularization is considered to smooth the final learning weights with reduced over-fitting [1,3]. In this study, learning rate is set 0.01, boosting algorithm is set to linear and tree, objective function is set to hinge. Stratified k-fold cross validation technique (k= 5) is used for reliable outcome with imbalance datasets. Four performance metrics namely precision (Pr), recall (Rc), F1-measure (F1), and area under the curve (AUC) are used to evaluate the performance of XGBoost [3].

3. Results and Discussion

The performance of XGBoost in discriminating the BBB drugs for linear and tree boosting is presented in Table I. Tree boosting approach obtained the highest performance of F1 = 78.5% . Both the tree and linear boosting yielded the highest recall of 94.3% for Ind features. XGBoost with tree classifiers yields highest AUC of 81.9% for combination of SE and Ind) features. Besides Tree method, XGBoost yielded the best AUC of 79.1% for combined features. For tree boosting method, XGBoost obtained higher than 75.0% Acc for SE, Ind and its combined features.

Table I: The performance of the XGBoost for linear and tree boosting methods

Metric	Tree			Linear		
	SE	Ind	SE + Ind	SE	Ind	SE + Ind
Pr	71.8	94.3	77.2	76.6	47.1	79.2
Rc	75.4	54.6	81.1	73.9	94.6	73.6
F1	72.9	67.2	78.5	74.2	62.6	75.7
AUC	77.3	75.5	81.9	78.4	58.7	79.1

4. Conclusion

The proposed methods are found to be capable of handling the multi variant features of drug identify BBB permeability.

References

[1] Miao R, Xia LY, Chen HH, Huang HH, Liang Y. Improved classification of blood-brain-barrier drugs using deep learning. Scientific reports. 2019 Jun 19;9(1):1-1.

[2] Doniger, S., Hofmann, T. Yeh, J. Predicting CNS Permeability of Drug Molecules: Comparison of Neural Network and Support Vector Machine Algorithms. Journal of Computational Biology 9, 849–864 (2002).

[3] Qiu Y, Zhou J, Khandelwal M, Yang H, Yang P, Li C. Performance evaluation of hybrid WOA-XGBoost, GWO-XGBoost and BO-XGBoost models to predict blast-induced ground vibration. Engineering with Computers. 2021 Apr 22:1-8.

Challenges of Trustable AI and Added-Value on Health
B. Séroussi et al. (Eds.)

doi:10.3233/SHTI220613

Performance of Machine Learning Methods to Classify French Medical Publications

Jamil ZAGHIR[a,b,1], Jean-Philippe GOLDMAN[a,b], Mina BJELOGRLIC[a,b], Daniel KESZTHELYI[a, b], Christophe GAUDET-BLAVIGNAC[a, b], Hugues TURBÉ[a, b], Belinda LOKAJ[a, b] and Christian LOVIS[a,b]
[a] *Division of Medical Information Sciences, University Hospitals of Geneva,*
[b] *Department of Radiology and Medical Informatics, University of Geneva, Switzerland*

Abstract. Many medical narratives are read by care professionals in their preferred language. These documents can be produced by organizations, authorities or national publishers. However, they are often hardly findable using the usual query engines based on English such as PubMed. This work explores the possibility to automatically categorize medical documents in French following an automatic Natural Language Processing pipeline. The pipeline is used to compare the performance of 6 different machine learning and deep neural network approaches on a large dataset of peer-reviewed weekly published Swiss medical journal in French covering major topics in medicine over the last 15 years. An accuracy of 96% was achieved for 5-topic classification and 81% for 20-topic classification.

Keywords. Document classification, unstructured medical data, machine learning, deep learning, natural language processing, Revue Médicale Suisse, French

1. Introduction

With the increasing amount of available data, information overload remains a challenge many organizations face. This is also true for healthcare research, where a large number of articles are published in medical journals. Seeking precise information in this motley collection of data is resource-demanding. As these articles cover the broad discipline of medicine, a way of helping information retrieval is to separate the dataset into multiple topics. If the collection size of data is too large to consider manual annotation, Automatic Document Classification (ADC) systems can be used to automatically assign labels. This research presents an approach for ADC, applying Natural Language Processing (NLP) methods coupled with Machine Learning pattern-identification abilities. A three-phase methodology to classify French free-text medical articles to their closest subject (i.e. multiclass classification) is proposed: (i) extract data to get documents and their corresponding topics; (ii) run NLP pipeline to preprocess documents, making them suitable for ADC methods; and (iii) classify documents using several approaches, from Traditional Machine Learning (TML) to Deep Learning (DL) models. To the best of our knowledge, no literature on document classification was published for French medical articles, the present paper aiming at filling this gap.

[1] Corresponding Author, Jamil ZAGHIR, Geneva, Switzerland; E-mail: jamil.zaghir@unige.ch.

2. Methods

The dataset contains more than 13 000 articles from the Revue Médicale Suisse [1], a weekly peer-reviewed medical magazine published, for the last 15 years, in French. The labeled set of data is built in two steps: (i) gathering textual content and the corresponding topics; and (ii) curation of the topics to reduce them from 880 to 298. Articles whose topics belong to the top-k of clinical specialties are kept, with k being 5 and 20 (6% and 17% of the dataset). For fair comparison, the text preprocessing step is identical for each learning method. The pipeline consists in converting text to lowercase, deleting punctuations, stop-words, and words occurring less than twice. Among models, DL classifiers have an embedding layer. One model computes the mean of word embeddings, and is based on a Feed-forward Network (FNN) while another one is a Convolutional Neural Networks (CNN) (4 layers, filters size 128). The train/test split is 80%/20%.

3. Results

The models' results are shown in Table 1 for 5 and 20 multiclass classification using TML and DL. FNN has the best results with an accuracy of 81% in top-20 classification.

Table 1. Classification performance with the following abbreviations P: Precision, R: Recall, Acc: Accuracy

	5-classes				20-classes			
Model	**P**	**R**	**F1-score**	**Acc**	**P**	**R**	**F1-score**	**Acc**
SVM (RBF)	0.92	0.91	0.91	0.91	0.73	0.71	0.71	0.71
SVM (Linear)	0.93	0.93	0.93	0.93	0.76	0.76	0.76	0.76
Naïve Bayes	0.92	0.92	0.92	0.92	0.75	0.75	0.74	0.74
Logistic Regression	0.94	0.93	0.94	0.94	0.78	0.78	0.78	0.78
Mean-Emb. + FNN	**0.96**	**0.95**	**0.96**	**0.96**	**0.83**	**0.81**	**0.81**	**0.81**
Emb. + 2D-CNN	0.92	0.91	0.91	0.92	0.74	0.71	0.70	0.70

4. Discussion

Even though FNN performed the best, Logistic Regression could be an interesting candidate as training is up to 10 times faster on this dataset for a performance of only 2 to 4% smaller accuracy. Since there are on average 151 documents per class for top-5 classification and 111 documents for top-20, CNN performed the worst. However, a higher amount of data might be needed, as in a similar work [2] showed CNN outperforming LR with 4000 samples per class. This research has been co-funded by "NCCR Evolving Language, Swiss National Science Foundation Agreement #51NF40_180888" and by "Leenaards Foundation".

References

[1] Revue Médicale Suisse - Revue médicale francophone de référence, (n.d.). https://www.revmed.ch/

[2] Hughes M, Li I, Kotoulas S, Suzumura T. Medical Text Classification Using Convolutional Neural Networks. Stud Health Technol Inform. 2017;235:246-250. PMID: 28423791.

Challenges of Trustable AI and Added-Value on Health
B. Séroussi et al. (Eds.)

doi:10.3233/SHTI220614

Analyzing the Information Content of Text-Based Files in Supplementary Materials of Biomedical Literature

Nona NADERI[a,b,1], Anaïs MOTTAZ[a,b], Douglas TEODORO[a,b,c], and Patrick RUCH[a,b]
[a] *University of Applied Sciences and Arts of Western Switzerland (HES-SO)*
[b] *Swiss Institute of Bioinformatics, Switzerland*
[c] *University of Geneva, Switzerland*

Abstract. We present an analysis of supplementary materials of PubMed Central (PMC) articles and show their importance in indexing and searching biomedical literature, in particular for the emerging genomic medicine field. On a subset of articles from PubMed Central, we use text mining methods to extract MeSH terms from abstracts, full texts, and text-based supplementary materials. We find that the recall of MeSH annotations increases by about 5.9 percentage points (+20% on relative percentage) when considering supplementary materials compared to using only abstracts. We further compare the supplementary material annotations with full-text annotations and we find out that the recall of MeSH terms increases by 1.5 percentage point (+3% on relative percentage). Additionally, we analyze genetic variant mentions in abstracts and full-texts and compare them with mentions found in supplementary text-based files. We find that the majority (about 99%) of variants are found in text-based supplementary files. In conclusion, we suggest that supplementary data should receive more attention from the information retrieval community, in particular in life and health sciences.

Keywords. Text mining, Semantic annotation, Supplementary materials.

1. Introduction

PubMed stores a wealth of supplementary materials; however, the analysis of such materials is mainly ignored in the literature. In this study, our aim is to examine text-based supplementary materials in terms of their semantic contents.

2. Methods

For our analysis, we have randomly selected 500 PMC articles that include text-based files such as spreadsheets in their supplementary materials. To estimate the importance of text-based supplementary files (spreadsheets, docx, and PDFs), we propose to evaluate MeSH terms extracted from them against MeSH terms manually assigned to reflect the manuscript content [3]. Since the goal is to assess the information content (not

[1] Corresponding Author, Nona Naderi, Information Science Department, HES-SO/HEG Geneva, Switzerland; E-mail: Nona.naderi@hesge.ch.

to achieve the state-of-the-art in information extraction), we use MetaMap text processing tool [1] to extract MeSH terms from abstracts, full-texts, and supplementary text-based files. We further examine these files to identify genetic variants. We use a set of regular expressions to retrieve the different formats of substitution variants. Our regular expressions capture standard formats, as defined by HGVS [2], non standard formats as found in the literature, as well as common variant database identifiers.

3. Results

The recall of MeSH terms using only abstracts is 22.2% and it increases to 28.1% by adding the annotations of 1,643 supplementary text-based files, that is a 5.9 percentage point increase in recall (+20% on relative percentage), which is statistically significant. We further analyze the annotations found in the full-texts and the recall of MeSH descriptors is 46.1%. By adding the annotations found in the supplementary text-based files, the recall increases to 47.6%, which is about a 1.5 percentage point increase (+3.3% on relative percentage). We have found 157 publications with at least one variant. Abstracts, full texts, and supplementary text files contain 14, 518, and 86,348 variants, respectively. The huge majority of variants are thus retrieved in supplementary materials, in particular in spreadsheets, representing 99% of variants. An expert has manually verified a set of 50 random variants found in the full-text, 50 random variants in the supplementary data, as well as the 14 variants found in the abstracts. The manual check of their validity confirms that we retrieve mainly real variants with our regular expressions: 100% of analyzed variants in abstracts and spreadsheets are correct. In the full-text, 78% are correct.

4. Conclusion

In this work, we have performed the analysis of the information content in text-based files in supplementary materials of PubMed articles using MeSH terms and variants. Our results suggest that supplementary materials should be considered as a source of information in the development of text mining tools for assigning MeSH terms to biomedical articles and indexing genetic variants in particular for personalized health.

Acknowledgement: This work has been supported by the CINECA project (H2020 No 825775).

References

[1] Aronson AR, Lang FM. An overview of MetaMap: historical perspective and recent advances. Journal of the American Medical Informatics Association. 2010;17(3):229-36.
[2] den Dunnen JT, Dalgleish R, Maglott DR, Hart RK, Greenblatt MS, McGowan-Jordan J, et al. HGVS recommendations for the description of sequence variants: 2016 update. 2016;37(6):564-9.
[3] Kans J. Entrez direct: E-utilities on the UNIX command line. National Center for Biotechnology Information (US); 2020.

Challenges of Trustable AI and Added-Value on Health
B. Séroussi et al. (Eds.)

doi:10.3233/SHTI220615

Named Entity Recognition in Pubmed Abstracts for Pharmacovigilance Using Deep Learning

T. Trang NGHIEM[a] and Cedric BOUSQUET[b,c,1]
[a]*Institute of Thermal, Mechanical and Material Sciences (ITheMM EA 7548), University of Reims Champagne-Ardenne, 51687 Reims, France.*
[b]*Unit of public health and medical informatics, CHU de Saint Etienne, France*
[c]*Sorbonne Université, Inserm, université Paris 13, Laboratoire d'informatique médicale et d'ingénierie des connaissances en e-santé, LIMICS, F-75006 Paris, France*

Abstract. Methods of natural language processing associated with machine learning or deep learning can support detection of adverse drug reactions in abstracts of case reports available on Pubmed. In 2012, Gurulingappa et al. proposed a training set for the recognition of named entities corresponding to drugs and adverse reactions on 3000 Pubmed abstracts. We implemented a classifier using deep learning with a Bi-LSTM and a CRF layer that achieves an F-measure of 87.8%. Perspectives consist in using BERT for improving the classifier, and applying it to a large number of Pubmed abstract to build a database of case reports available in the literature.

Keywords. Artificial intelligence, pharmacovigilance, Pubmed, Deep learning

1. Introduction

In order to monitor drug safety with the medical literature, it is necessary to carry out bibliographic queries on a regular basis. However, pharmacovigilance teams have only Pubmed and its user interface which is intended for all users. Our aim is to build a database consisting of Pubmed abstracts related to adverse drug reactions (ADR). We employ the training set based on 3000 Pubmed abstracts on case reports likely to describe ADRs implemented by Gurulingappa et al. [1]. Most previous works have applied machine-learning-based methods, but not deep learning [2, 3].

Our first step as described in this paper was to implement a classifier to detect named entities related to drugs and ADRs using a conditional random field (CRF) and a Long short-term memory (LSTM) deep learning layers. Our model can be trained quickly (less than 1 hours) on only one 16 GB RAM computer. We compared our results with those of other recent works using named entity recognition of drugs and ADRs with deep learning [4, 5, 6].

[1] Corresponding Author, Dr Cedric Bousquet, SSPIM, Bâtiment CIM42, chemin de la Marandière, Hôpital Nord, 42055 Saint Etienne; E-mail: cedric.bousquet@chu-st-etienne.fr.

2. Methods

We used a set of 3000 Pubmed abstracts annotated in 2012 by Gurulingappa *et al.* that contained 6821 sentences with a relation between a drug and an adverse reaction, and 16,695 sentences without [1]. We implemented a CRF with sklearn-crfsuite and two versions of a Bi-LSTM deep learning model with or without a CRF layer. Embeddings of size 40 were fed into the network using Keras' embedding layer. We used an IO tagging of tokens in Pubmed abstracts where "I" corresponds to "Inside", including two types of "I" (I_ADR and I_DRUG), and "O" stands for "Out". Each experiment was repeated five rounds, and the averaged results were taken for all metrics (precision, recall and F1-score). Obtained results are reported below with those of previous publications.

3. Results and Discussion

Table 1 shows the results we obtained with our models (first three lines), and results obtained by other authors (next three lines).

Table 1. Precision, recall and F1-Score for our models and other publications. POS stands for Part-of-speech in Named Entity Recognition model.

Models	Precision (%)	Recall (%)	F1-Score
CRF (with POS)	92.8	81.1	86.6
Bi-LSTM	85.7	85.6	85.6
Bi-LSTM + CRF	92.6	83.5	87.8
Ramamoorthy et al. [4]	88.4	82.4	85.3
P. Ding et al. [5]	86.7	94.8	90.6
Hussain et al. [6]	98.2	96.4	97.6

Our lightweight model is capable to extract quite efficiently named entities on drugs and adverse reactions from Pubmed abstracts. We plan to detect ADRs on a large number of case reports to build a reference database to improve queries for pharmacovigilance in the literature. A Bidirectional Encoder Representations from Transformers (BERT) model [7] should be implemented to improve the performances of our classifier.

References

[1] Gurulingappa, H, Mateen-Rajpu A, Toldo L. Extraction of potential adverse drug events from medical case reports. J biomedical semantics. 2012;3(1):1-10.
[2] Sarker A, Ginn R, Nikfarjam A, O'Connor K, Smith K, Jayaraman S, Upadhaya T, Gonzalez G. Utilizing social media data for pharmacovigilance: A review. J Biomed Inform. 2015;54:202-12.
[3] Luo Y, Thompson WK, Herr TM, Zeng Z, Berendsen MA, Jonnalagadda SR, Carson MB, Starren J. Natural Language Processing for EHR-Based Pharmacovigilance: A Structured Review. Drug Saf. 2017;40(11):1075-1089.
[4] Ramamoorthy S, Murugan S. An attentive sequence model for adverse drug event extraction from biomedical text. arXiv preprint. 2018. arXiv:1801.00625.
[5] Ding P, Zhou X, Zhang X, Wang J, Lei, Z. An attentive neural sequence labeling model for adverse drug reactions mentions extraction. IEEE Access. 2018; 6:73305-15. doi: 10.1109/ACCESS.2018.2882443
[6] Hussain S, Afzal H, Saeed R, Iltaf N, Umair MY. Pharmacovigilance with Transformers: A Framework to Detect Adverse Drug Reactions Using BERT Fine-Tuned with FARM. Comput Math Methods Med. 2021;2021:5589829. doi: 10.1155/2021/5589829. eCollection 2021.
[7] Devlin J, Chang MW, Lee K, Toutanova K. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint. 2018. arXiv:1810.04805.

Challenges of Trustable AI and Added-Value on Health
B. Séroussi et al. (Eds.)

doi:10.3233/SHTI220616

Deep Learning Methods for Detecting Side Effects of Cancer Chemotherapies Reported in a Remote Monitoring Web Application

Marie-Hélène METZGER[a,b,1], Ahmath GADJI[a], Nada HAJ SALAH[a], Wedji KANE[a] and François BOUE[b]
[a] *Equipe soins primaires et prévention, INSERM U1018, Villejuif, France*
[b] *Service de médecine interne, Hôpital Antoine-Béclère, Assistance Publique Hôpitaux de Paris, Clamart, France*

Abstract. The objective of our work was to develop deep learning methods for extracting and normalizing patient-reported free-text side effects in a cancer chemotherapy side effect remote monitoring web application. The F-measure was 0.79 for the medical concept extraction model and 0.85 for the negation extraction model (Bi-LSTM-CRF). The next step was the normalization. Of the 1040 unique concepts in the dataset, 62, 3% scored 1 (corresponding to a perfect match with an UMLS CUI). These methods need to be improved to allow their integration into home telemonitoring devices for automatic notification of the hospital oncologists.

Keywords. Telemedecine, Antineoplastic agents, Drug-related Side Effects and Adverse Reactions, Natural Language Processing, Deep Learning

1. Introduction

Cancer chemotherapies are highly toxic and some of them can cause serious side effects, requiring urgent hospital treatment. In a randomized trial evaluating the use of a remote monitoring web application of cancer chemotherapies, patients reported their side effects daily. The objective of our work was to use deep learning methods for extracting and normalizing side effects reported in free text by patients in the web application.

2. Methods

Study design: Patients were recruited from the Antoine-Béclère hospital located in the southern surburbs of Paris (Assistance Publique – Hôpitaux de Paris). The NEUTROSIS project is a research project that consisted of an evaluation of the impact on hospital management of a daily home monitoring web application for patients undergoing cancer chemotherapy. The study was designed as a randomized controlled trial. One arm used

[1] Corresponding Author, Service de médecine interne, Hôpital Antoine-Béclère, Assistance Publique Hôpitaux de Paris, 157 rue de la porte de Trivaux, Clamart, France; E-mail: metzger.marie-helene@orange.fr.

the web application (n=49) allowing the patient to enter his temperature and symptoms each morning for 14 days following each chemotherapy cycle. The control group (n=51) did the same monitoring using a paper diary. Patients reported several symptoms in structured form (temperature, diarrhea, vomiting, mucositis) but could also report other symptoms in free text. The patients were monitored during a period of six months while undergoing chemotherapy treatment, from 24/02/2017 to 17/05/2019.

Extraction and normalization processing of symptoms related to side effects of cancer chemotherapies: The extraction method was based on deep learning with word embeddings using bidirectional long short-term memory network-conditional random field (Bi-LSTM-CRF) [1]. One model was developed for medical concepts extraction and one model for negation extraction. The two models were applied in series. The performance of the models was evaluated by calculating F-measure of the extraction models against a referential (manual annotation of NEUTROSIS text data by a physician).

After extraction, the next step was to standardize the medical concepts using the Unified Medical Language System (UMLS). The first normalization step consisted in searching the medical concepts in the French version using the SimString tool to match the medical concepts with the UMLS [2]. The second step consisted of the use of the UMLSBertALL model [3]. Standardization was assessed manually by medical review of the medical concept and its CUI correspondence. A score of 1 was assigned if there was a perfect match, allowing a computation of the recall.

3. Results

On the 12,665 annotated tokens, the F-measure was 0.79 for the medical concept extraction model and 0.85 for the negation extraction model. Of the 1,040 unique concepts of the set, applying a similarity threshold of 0.9 for the first step and a threshold of 0.4 for the second step of normalization, 62, 3% of the concepts obtained a score of 1 (recall).

4. Discussion and Conclusions

The performance of applied methods for extracting and normalizing chemotherapy side effects from patients is not high enough to allow their application in remote monitoring devices. The next step will be to apply active learning methods to re-train the models.

References

[1] Habibi M, Weber L, Neves M, Wiegandt DL, Leser U. Deep learning with word embeddings improves biomedical named entity recognition. Bioinformatics (Oxford, England). 2017;33:i37–48.

[2] Okazaki N, Tsujii J. Simple and Efficient Algorithm for Approximate Dictionary Matching. In: Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010). Beijing, China: Coling 2010 Organizing Committee; 2010. p. 851–9

[3] Yuan Z, Zhao Z, Sun H, Li J, Wang F, Yu S. CODER: Knowledge-infused cross-lingual medical term embedding for term normalization. J Biomed Inform. 2022;126:103983.

This page intentionally left blank

Section VIII

Security and Safety; Telehealth; Sensors, Signals and Imaging Informatics

This page intentionally left blank

Challenges of Trustable AI and Added-Value on Health
B. Séroussi et al. (Eds.)

doi:10.3233/SHTI220618

Patient Safety Informatics: Criteria Development for Assessing the Maturity of Digital Patient Safety in Hospitals

Jan-Oliver KUTZA[a], Ursula HÜBNER[a,1,] A Jay HOLMGREN[b] and Moritz ESDAR[a]
[a] *Health Informatics Research Group, Osnabrück University AS, Germany*
[b] *Center for Clinical Informatics and Improvement Research, University of California San Francisco, USA*

Abstract. With the start of the 21st century, patient safety as a topic of special interest has attracted increasing attention in both academia and clinical practice. As technology has continued to develop since then, questions and focal points surrounding the topic have also shifted. In particular, questions regarding the impact of digitalization on patient safety and its measurement are now of high interest. This work aims to develop a maturity assessment instrument in the form of a criteria set for measuring structural requirements for digital patient safety in hospitals. Based on the results of a literature review and a derivation of maturity objects (MO) from known maturity models, 64 criteria across 11 categories were developed. Written comments of two digital patient safety experts as well as subsequent interviews were used to evaluate and refine the criteria catalog. The resulting catalog offers hospitals guidance for detecting possible areas of structural improvements in their information systems with regard to patient safety and represents a unique instrument for assessing digital maturity in this particular area.

Keywords. Patient Safety, Hospital Information Systems, Digital Maturity Models

1. Introduction

"To Err is Human", published by Kohn and colleagues in 2000, marked the beginning of renewed focus by healthcare providers and researchers to find ways of ensuring and improving patient safety [1, 2]. The report highlighted serious problems related to a high level of medical errors and pointed to potential solutions, including calls for better information systems. However, two decades later, low patient safety continues to be a pressing concern, and many developments in this domain have happened unexpectedly slowly [3]. At the same time, healthcare providers have continued to digitize their care delivery processes and information systems [4]. While digitalization is generally expected to have positive effects on both costs and quality of care [5], it can also induce various unintended consequences [6]. It may solve old familiar problems in terms of medication errors by introducing clinical decision support that intervenes at the point of care, but also has the potential to create new ones [6, 7]. Unintended consequences such as alert fatigue or adverse effects such as misidentifications of patients can arise when

[1] Corresponding Author, Ursula Hübner, Osnabrück University of AS, Health Informatics Research Group, PO Box 1940, 49009 Osnabrück, Germany, E-mail: u.huebner@hs-osnabrueck.de.

the requirements for health IT (HIT) systems are ill-defined, or as clinicians take time to adapt to new workflows and processes.

It is therefore of utmost importance for policy makers and hospitals to not just blindly push for higher degrees of digitalization for its own sake but instead work against a set of goal-orientated criteria that increase the likelihood of reducing medical errors. Despite the unequivocal importance and potential of ensuring patient safety by means of well-designed HIT systems, no standardized measures exist that gauge their maturity regarding their patient safety performance in particular. Existing maturity models such as the Electronic Medical Record Adoption Model (EMRAM) tend to focus on the availability of various, rather generic IT functions (such as nursing documentation) as well as the order in which they are implemented. These models are criticized, among other things, for invoking misguided incentives by enforcing a predetermined path to digitalization [8].

To provide comprehensible guidance as to how patient safety can be ensured when digitalizing hospital care processes, this research aims to develop a set of criteria that can point health care organizations such as hospitals, policymakers and vendors towards creating a safer environment for patients.

2. Methods

A three-stage development process comprising the following steps was carried out: First, we conducted a literature search in the databases Pubmed Medline, Wise, CINAHL, Scinos, and Google Scholar. The search sought to identify publications that were related to patient safety in the form of a condition, trigger, adverse event, or outcome. The combination of the two search terms "patient safety" AND "hospital" was linked with each of the following terms "digit*", "high risk company", "risk management", "management of adverse events", "adverse event", "error", "maturity model", "German Hospital Future Act[2](KHZG)[3]". The search was performed between March and August of 2021 and was limited to English and German publications from 2010 to 2021.

In the second step, five maturity models (CHECK-IT [9], EMRAM [10], WCS [11], KIT-CON [12], and MOST WIRED [13]) as well as the KHZG's eligibility criteria were reviewed for maturity objects (MOs). They are criteria for assessing digital maturity (e.g. "vital signs monitoring of emergency patients using medical devices") [12]. MOs that were unrelated to patient safety were removed. Based on the topical structure of the German Hospital Future Act, a categorial system, consisting of categories like “digital medication management” or “digital documentation” was developed, and the identified MOs were then assigned to these categories. This procedure was repeated for patient safety items (PSI), like “time pressure/ high workload” or “incomplete communication”, that were extracted from the literature review. The PSI and MOs were logically assigned to each other. If MOs could not be assigned to a PSI and vice versa, they were excluded from further use. For example, the PSI “informal standards” could not be reasonably matched to a corresponding MO and was therefore excluded. The criteria were then formed from the remaining MOs and adapted in wording according to [14, 15].

At last, two experts (physicians with a background in patient safety work and health informatics) were asked to comment on the criteria for validation purposes. The

[2] Act for funding the digitization process in German hospitals with a total volume of 4.3 billion Euros

[3] In the actual search, the German word “Krankenhauszukunftsgesetz” (KHZG) was used.

comments were then discussed in individual digital interviews (average duration 20 minutes). The criteria were subsequently updated by the authors (JOK, AJH, ME) and adjusted where necessary.

3. Results

The literature search yielded a total of 85 relevant articles that contained information about factors for HIT, patient safety or a combination of both. 43 patient safety items (PSI) could be derived from the literature. From the maturity models, 2350 maturity objects (MO) were extracted, which were reduced to 78 unique MOs related to patient safety over three iterations. Twelve categories were initially derived from the KHZG. After assigning the MOs and PSI to the categories and examining overlaps, 65 criteria (divided between eleven categories) remained which were used for the development of the criteria catalog. Following the interviews, the combination of three criteria was found to be necessary as well as the inclusion of a new one, resulting in the final catalog (Tab.1) that comprises a total of 64 criteria in eleven categories.

Table 1. Exemplary excerpt from the criteria catalog for assessing digital patient safety maturity in hospitals. The full catalog can be found here: https://1drv.ms/b/s!ApxR0mqhcuA1jf47WtS4xJOEcCfiVQ?e=YrYESo

Category	Criteria
Digital documentation	Digital documentation is standardized throughout the organization in terms of nomenclature, coding and form.
	All digital inputs are legible, clear and unambiguous for the user.
Digital file	All patient data of the current treatment as well as data of previous and external treatments (if provided by the patients), are available in the digital patient file.
	In the event of a system failure, it is ensured that health professionals continue to have full access to important patient data (allergies, problems, diagnoses, medication, lab results, progress logs, vital signs) at the point of care.
Digital medication management	The entire medication management process is carried out digitally in the form of closed-loop medication management (ordering, documentation, testing, positioning, administration, etc.).
	Systemic testing for medication errors, over/under doses, drug allergies, contraindications, drug-food interactions is possible and can generate alerts.
Digital treatment management	Pathologically deviating vital signs generate automatic alerts.
	The execution of complex and standardized activities for the treatment of patients (e.g. operations, hygiene measures, mobilization, etc.) is checked by medical professionals for complete and correct execution using digital checklists.
Digital discharge management	All information items provided to patients at the time of discharge is legible, unambiguous, correct, complete, and available to them in standardized digital and machine-readable form for all common platforms, as well as in paper form if required.
	A digital based assessment of discharge risk is performed, and alerts are issued in the event of increased risk potential for the patient.
Digital decision support system/ risk assessment system	The digital treatment and documentation software identifies potential risks based on the complete patient data (falls, pressure ulcers, multi resistant pathogens, nosocomial infections, malnutrition, pain, incontinence, injuries, death, etc.) and generates alerts to make health professionals aware of them.
	An automatic review of all prescriptions of clinical relevance is performed and recommendations regarding potential alternatives are issued. This is also done when patients are admitted with their existing orders (such as home medication).
Digital service request	Service requests are made digitally only and generate notifications for the service provider.

	Automatic warning messages are generated as soon as a service request is created twice, which are only requested once for comparable treatment processes (e.g. dialysis once a day vs. twice a day).
Robotics, hardware and software	The preparation of individual doses (medication) is robot/dispensing-machine controlled.
	Medical equipment used on patients without continuous supervision by health professionals (e.g. monitoring equipment) is connected and continuously as well as automatically checked for error messages or serious deviations.
Digital incident and error management	For patterns involving nears-miss events a continuous clinic-wide background check is performed, which is compared with near-miss events, errors and incidents of harm in order to issue appropriate warnings about possible correlations of causal chains.
	Employees have the opportunity to immediately report errors and undesired events digitally and anonymously.
Digital patient observation	Digital observation of patients can be performed remotely outside the point of care (e.g., by an on-call physician).
	Locating patients with a tendency to wander in or out (e.g., in the case of dementia) is digitally possible in certain areas (e.g., geriatrics).
Digital information transfer	The transmission of patient data between the actors involved in the treatment (external) takes place exclusively digitally.
	A uniform patient identification number is used in cross-sector and cross-organizational communication.

4. Discussion

Patient safety continues to be one of the key issues for providing high quality patient care. While digitalization can be used to reduce barriers for achieving maximum patient safety, it can also create hurdles if used inappropriately and missing adequate control. This study provides a categorial system and related criteria that is based on the international literature, existing maturity models and expert discussions. It provides a framework to evaluate the maturity of health IT systems regarding their compliance with structural patient safety requirements.

While previous studies on patient safety primarily focused on medication, this study includes the patient journey within a hospital covering any type to treatment, patient monitoring and transfer and discharge management. It also refers to accountability of patient safety measures in terms of documentation, incident management and (electronic) availability of information. In summary, it provides more details than other approaches and abandons the rationale of mere rank order of IT systems availability such as EMRAM.

However, some limitations have to be considered. Although, this study is based on 5 maturity models, it is recommended to include further models in future research. Also, the use of a systematic literature review could generate further results regarding patient safety items.

Nevertheless, this research is the first to provide a comprehensive set of practical requirements that must be placed on HIT systems to meet the promise of improved patient safety specifically. It could be used as an easily accessible list of items that hospitals can use to review their internal digital processes and structures to identify potential patient safety threats in advance. Additionally, it could be used as a basis for designing maturity assessment for policymakers on a national and international level. Being the first of its kind, the criteria catalog has yet to be tested and evaluated in practice. Future research could focus on validation and evaluation as well as the definition of an appropriate scoring scheme.

Acknowledgement

We would like to thank Thorsten Kaiser (PD Dr. med.), laboratory physician at the Institute for Laboratory Medicine, Clinical Chemistry and Molecular Diagnostics University Hospital Leipzig as well as Reinhard Strametz (Prof. Dr. med. Dipl.-Kfm.), professor at RheinMain University of Applied Sciences for assessing the criteria.

References

[1] Kohn LT, editor. To err is human: Building a safer health system. 7th ed. Washington, DC: National Acad. Press; 2000. 287 p. eng.
[2] Stelfox HT, Palmisani S, Scurlock C, Orav EJ, Bates DW. The "To Err is Human" report and the patient safety literature. In: p. 174–8. (vol. 15). eng.
[3] Lark ME, Kirkpatrick K, Chung KC. Patient Safety Movement: History and Future Directions. In: p. 174–8. (vol. 43). eng.
[4] Guo C, Ashrafian H, Ghafur S, Fontana G, Gardner C, Prime M. Challenges for the evaluation of digital health solutions-A call for innovative evidence generation approaches. NPJ Digit Med. 2020;3110.
[5] Keasberry J, Scott IA, Sullivan C, Staib A, Ashby R. Going digital: a narrative overview of the clinical and organisational impacts of eHealth technologies in hospital practice. Aust Health Rev. 2017;41(6):646–64. doi:10.1071/AH16233 Cited in: PubMed; PMID 28063462.
[6] Colicchio TK, Cimino JJ, Del Fiol G. Unintended Consequences of Nationwide Electronic Health Record Adoption: Challenges and Opportunities in the Post-Meaningful Use Era. J Med Internet Res. 2019;21(6):e13313. doi:10.2196/13313 Cited in: PubMed; PMID 31162125.
[7] Schneider E, Ridgely S, Meeker D, Hunter L, Khodyakov D, Rudin R. Promoting Patient Safety Through Effective Health Information Technology Risk Management: RAND Corporation; 2014.
[8] Cresswell K, Sheikh A, Krasuska M, Heeney C, Franklin BD, Lane W, Mozaffar H, Mason K, Eason S, Hinder S, Potts HWW, Williams R. Reconceptualising the digital maturity of health systems. In: e200-e201. (vol. 1).
[9] Marburger Bund. Check IT - Das Marburger Bund-Analysetool zum digitalen Krankenhaus | Marburger Bund [Internet]. 2022 [updated 2022 Jan 24; cited 2022 Jan 24]. Available from: https://www.marburger-bund.de/bundesverband/themen/krankenhauspolitik/digitale-modernisierung/check-it-das-marburger-bund
[10] HIMSS Analytics - Europe. Electronic Medical Record Adoption Model [Internet]. 2017 [updated 2021 Jan 12; cited 2022 Jan 24]. Available from: https://www.himssanalytics.org/europe/electronic-medical-record-adoption-model
[11] Esdar M, Hübner U, Thye J, Babitsch B, Liebe J-D. The Effect of Innovation Capabilities of Health Care Organizations on the Quality of Health Information Technology: Model Development With Cross-sectional Data. In: e23306. (vol. 9). eng.
[12] Dickmann F, Kümmel K, Oroszi F, Schneider M, Rienhoff O, editors. Der IT-Reifegrad von Krankenhäusern: Edition 2020. Glückstadt: Verlag Werner Hülsbusch Fachverlag für Medientechnik und -wirtschaft; 2020. 190 p. ger.
[13] Healthcare IT - CHIME. Digital Health Most Wired - Healthcare IT - CHIME [Internet]. 2021 [updated 2021 Dec 8; cited 2022 Jan 24]. Available from: https://chimecentral.org/chime-most-wired-2/#tab_ert_pane1-0
[14] Neff AA, Hamel F, Herz TP, Uebernickel F, Brenner W, vom Brocke J. Developing a maturity model for service systems in heavy equipment manufacturing enterprises. Information & Management. 2014;51(7):895–911.
[15] Bareuther V. Entwicklung eines spitalbezogenen eHealth Reifegradmodells: und Validierung am Beispiel Hirsladen [Masterarbeit]. Zürich: Hochschule für angewandte Wissenschaften; 07.07.2017.

Challenges of Trustable AI and Added-Value on Health
B. Séroussi et al. (Eds.)

doi:10.3233/SHTI220619

The First Introduction of Social Robotics in Rehabilitation Care

Marian HURMUZ[a,b,1], Stephanie JANSEN-KOSTERINK[a,b], Ina FLIERMAN[c], Behrouz FARD[c] and Lex van VELSEN[a,b]
[a] *eHealth department, Roessingh Research and Development, The Netherlands*
[b] *Biomedical Signals and Systems Group, University of Twente, The Netherlands*
[c] *Roessingh Center for Rehabilitation, The Netherlands*

Abstract. Can you imagine to receive treatment through a robot? When talking about the future of healthcare, this is the vision many people have. Currently, the predominant role of social robots in care is entertaining patients. However, this does not have an impact on care process itself. In this paper, we focus on defining use cases other than merely keeping patients' company by implementing a Pepper robot in inpatient rehabilitation setting, and expand upon usability testing the use cases. Our findings showed that, to ensure sustainable implementation of social robots in care organizations, we need excessive collaboration with the target population.

Keywords. Social robots, rehabilitation care, use cases, usability

1. Introduction

"It is 2040, Hannah (an inpatient patient) has an appointment planned today with Robin (a social robot). Robin helps Hannah with treating her chronic pain." When people think about the future of healthcare organisations, it often includes robotics [1]. Looking at their current use with healthcare systems, we see that there is a limited availability of social robots with actual on-site tasks and responsibilities. A well-known example of a social robot is Tessa. Tessa is a small flower pot which can, for instance, remind older adults of appointments, meals, daily activities [2].

However, most social robots in healthcare are currently used for entertaining patients and keeping them company [3]. In this kind of use, the added value of the social robot will not be directly linked to the core activity of the organisation in any meaningful way. Though social robots may provide patients with some entertainment and company, there may be much to be gained by, despite aiming to entertain patients and have social interaction with them, also aiming to relieve the workload of healthcare professionals (HCPs). Rehabilitation care has been recognized as a promising setting for the application of social robots [4]. Within the SCOTTY project, we will study whether rehabilitation care is indeed a promising setting for social robots. We used the co-design method to define use cases in conjunction with the target population. The aim of this paper is to describe the use cases defined for a social robot in inpatient care setting, and to show the first results of the usability testing of its implementation.

[1] Corresponding Author, Marian Hurmuz, eHealth department, Roessingh Research and Development, Roessinghsbleekweg 33b, 7522 AH, Enschede, The Netherlands; E-mail: m.hurmuz@rrd.nl.

2. Methods

In the SCOTTY project (DIH-HERO TTE, grant No 825003), we will implement the Pepper robot [5] in a rehabilitation centre. In this paper, we present our methods and results in two parts: the development of the use cases and its usability testing.

2.1. Use cases

For defining the use cases for the Pepper robot, we started with identifying the views of persons admitted for inpatient rehabilitation in Roessingh Center for Rehabilitation (RCR), and HCPs working at RCR about the Pepper robot through questionnaires. With these questionnaires, we identified the attitude towards robots, the intention to use the robot, and the tasks the robot could perform. The results of these questionnaires were analyzed and were the main input for two co-creation sessions to define the use cases. The first session focused on establishing the general functionalities. The second session focused on the roles of Pepper and the use cases. As rehabilitation medicine is a multidisciplinary field, a diverse group participated in the co-creation sessions: nurse, nurse in training, rehabilitation physician, innovation manager and researchers of the SCOTTY project. Finally, after both co-creation sessions, the outcomes were shared with the technical developers within the project to implement the use cases.

2.2. Usability study

After implementing the use cases within the robot, a usability study was conducted among technical experts, nurses (in training) and patients. This study did not require formal medical ethical approval (CMO Arnhem-Nijmegen file number: 2021-12988). The main outcomes were: usability performance metrics (i.e. task completion rate, time and satisfaction) and usability issues. Participants received pre-defined tasks to perform with Pepper. Patients received other tasks than the experts and HCPs. The tasks of experts and HCPs were: (1) Open the Scotty application, (2) Sign in, (3) Synchronize the newest agenda in the calendar, (4) Complete the vital functions questionnaire, and (5) Complete the fluid balance questionnaire. The tasks for the patients were: (1) Complete the USER-P self-report questionnaire, (2) Perform the following physical exercise: 'hip stretching in prone position', and (3) Play the solitaire card game for 1 minute.

During the usability tests, the think-aloud procedure was used (i.e. participants were encouraged to share their thoughts). The usability tests were voice- and video-recorded to gather the usability issues. These recordings were transcribed and analyzed. Furthermore, during the tests the researcher took notes of two of the usability performance metrics: task completion and time. After performing each task, participants completed a questionnaire to assess the third performance metric: task satisfaction.

3. Results

3.1. Use cases

A total of 13 spinal cord injury patients and 23 HCPs completed the questionnaire. In both groups, more females participated (54% of the patients and 87% of the HCPs), with

a mean age of 61.3 (SD=17.8) in the patient group and 38.7 (SD=13.5) in the HCP group. Both groups had a positive attitude towards using robots and a positive intention to use robots. On a scale from 1 (negative) to 5 (positive), patients' attitude was scored with a mean of 3.7 (SD=0.8, range=2–5), and HCPs' attitude with 3.7 (SD=0.6, range=2.3–5). On the same scale, patients' intention was scored with a mean of 3.8 (SD=1.0, range=2–5), and HCPs' intention with 3.6 (SD=0.7, range=2–4.7).

Patients were most positive about completing questionnaires an playing games with the Pepper robot. HCPs were most positive about using the Pepper robot for playing games with patients. Table 1 shows these results.

Table 1. Percentages patients and HCPs positive about conducting particular tasks with Pepper robot.

Tasks Pepper	% patients positive	% HCPs positive
Completing questionnaires	85	70
Conducting physical exercises	54	74
Playing games	85	92

During the two co-creation sessions, the outcomes of the questionnaires were discussed. Based on these sessions and the technical feasibility, the co-creation group (HPCs and lead researchers) formulated four potential roles for the Pepper robot. For each of these roles, different use cases were defined. The roles and final use cases are shown in Table 2.

Table 2. Use cases defined for Pepper for four different roles, based on questionnaires and co-creation sessions.

Nurse's aid	Physical therapist's assistant	Companion	Host
To note and store routine vital signs (e.g. temperature, blood pressure, pulse rate etc.) (	To remind patients to perform their routine exercises	To play a game with the patient	To provide general information of the health care facility and care process
Facilitate self-report routine questionnaires and store outcomes	To show patients their training videos	To read a book with the patient	To perform simple evaluation questionnaires

3.2. Usability study

A total of 12 adults participated in the usability study. Four experts participated, of which three were female and their age range was 20-27. Three nurses participated, all female aging from 19 to 50. Five patients participated (3 males, 2 females) with an age range of 19 to 77 years.

Looking at the different tasks the participants had to complete, the tasks considering completing questionnaires, were difficult to complete for all roles. Furthermore, the experts had trouble with completing the sign in task. All five experts used the wrong sign in card for this. See Table 3 for an overview of the usability metrics.

Table 3. Usability metrics of each task divided into the three roles: technical expert, nurse, patient.

Role	Task	% (rate) task completion	Range task completion time	Range task satisfaction*
Technical expert	1: Open Scotty app	75 (3/4)	8 – 10 sec	3.0 – 6.3
	2: Sign in	0 (0/4)	X	5 – 5.3
	3: Synchronize agenda	100 (4/4)	15 – 21 sec	5.3 – 6.0
	4: Vital functions	50 (2/4)	179 – 228 sec	1.3 – 5.7

	5: Fluid balance	75 (3/4)	147 – 200 sec	4.3 – 6.0
Nurse	1: Open Scotty app	100 (3/3)	9 – 12 sec	5 – 6.7
	2: Sign in	100 (3/3)	19 – 69 sec	4 – 4.7
	3: Synchronize agenda	100 (3/3)	12 – 20 sec	5.3 – 7.0
	4: Vital functions	33.3 (1/3)	166 sec**	1.0 – 6.0
	5: Fluid balance	66.7 (2/3)	141 – 163 sec	1.0 – 6.7
Patient	1: USER-P	0 (0/5)	X	1.0 – 3.3
	2: Physical exercise	100 (5/5)	75 – 248 sec	3.0 – 7.0
	3: Solitaire game	100 (5/5)	77 – 100 sec	4.7 – 7.0

*Task satisfaction measured on a scale from 1 (not satisfied) to 7 (satisfied)
** N=1, so only one task completion time

Regarding the usability issues, the nurses experienced most issues (N=17), followed by patients (N=15) and experts (N=14). Table 4 shows the number of issues found among the three groups divided into severity categories. We focus only on the critical issues, as these are the ones that have to be solved before implementing the Pepper robot.

Table 4. Number of usability issues (divided into three severity categories: minor, serious, critical) per role.

Usability issues severity	N issues among technical experts	N issues among nurses	N issues among patients	N issues among all roles
Minor	6	3	6	15
Serious	4	10	5	19
Critical	4	4	4	12
Total	14	17	15	53

Among the experts and nurses the same critical issues were identified. Regarding the critical issues identified by patients, one issue is shared by nurses and experts. That issue is that when completing questionnaires, it is unclear what kind of answer the Pepper robot is looking for. When users need to sign in (task 2 for experts and nurses), they have trouble finding the right button to open the sign in page. The icon used for that, was not recognized as a sign in button, except for one expert. Furthermore, when wanting to open a questionnaire, the QR code from the patient needs to be scanned. But among both experts and nurses it was unclear which QR code was the patient's. This critical issue was observed in both task 4 and 5. The last critical issue identified by experts and nurses for task 4 and 5, is that it was unclear whether the user had to wait with giving his/her answer to the Pepper robot until its blue lights turn on. This was frustrating for them as they had to repeat themselves multiple times. The three remaining critical issues that were identified only by patients and occurred all in task 1. When opening a questionnaire, they had to scan their QR code, but (1) it was unclear that they had to scan something, and (2) if they knew they had to scan their QR code, they did not know how to scan the code. Finally, when answering the questionnaire, it was unclear whether they had to give the answers out loud, or need to type in the response. This last critical issue did not appear among all patients – bear in mind that not all patients could even open the questionnaire – but among the ones where it did, it took too much time to be able to complete the task.

4. Discussion

We defined four roles for the robot, each with use cases. By actively involving the target population, developers can program a robot that better fits the end-users' needs.

After reaching an agreement on the use cases for Pepper within the SCOTTY project, the usability tests showed us that there are some critical issues that need to be solved. Conducting usability tests with social robots is a prerequisite before implementing such a device in a rehabilitation centre. With usability tests, we can assess our preconceived assumptions in practice, eliminate (unforeseen) errors and improve users' satisfaction with the system. When a system has too many errors, it irritates users, and users will discontinue their use [6,7]. The usability issues we encountered in our study will be solved by improving the technology and by educating users before implementation.

In conclusion, we propose that by using co-design for use case development for a social robot involving the target population, and by usability testing the robot among the target population, the implementation of such a robot will experience less difficulties. Of course, it is essential to keep evaluating the robot during the implementation stage and to keep improving the robot to reach sustainable implementation. After months of developing use cases, implementing them in the Pepper robot and testing the robot in the SCOTTY project, we have now arrived to the stage of final improvement and actual implementation in RCR. During this stage, we will continue to monitor the feasibility and added value of our social robot as experienced by HPCs and patients.

Acknowledgements

We want to thank the team of Bluecompanion for their work with the robot, Lisa Boerrigter and Marleen Nijenhuis for their help with developing use cases, and students from Saxion's Smart Solutions Semester for their help with the usability tests. This study was supported by the SCOTTY project (DIH-HERO technology transfer experiment), funded by EU's Horizon 2020 research and innovation program (No 825003).

References

[1] Davies N. Can robots handle your healthcare? Eng Technol. 2016 Oct;11(9):58–61.
[2] Vilans. Maatschappelijke Businesscase Dagstructuurrobots. 2020 [accessed 2022 Jan 3]. Available from: https://assets.website-files.com/5f96c0f3619f62286b45df33/600952ce4b478b4b1038f67b_maatschappelijke-businesscase-dagstructuurrobots.pdf.
[3] Aymerich-Franch L, Ferrer I. Socially assistive robots' deployment in healthcare settings: a global perspective. arXiv Prepr. 2021;arXiv:2110.07404.
[4] Wolbring G, Yumakulov S. Social Robots: Views of Staff of a Disability Service Organization. Int J Soc Robot. 2014 Mar;6(3):457–68.
[5] Softbank Robotics. Pepper the humanoid and programmable robot. [accessed 2022 Jan 6]. Available from: https://www.softbankrobotics.com/emea/en/pepper.
[6] Marien S, Legrand D, Ramdoyal R, Nsenga J, Ospina G, et al. A User-Centered design and usability testing of a web-based medication reconciliation application integrated in an eHealth network. Int J Med Inform. 2019 Jun;126:138–46.
[7] Wu D, Moody GD, Zhang J, Lowry PB. Effects of the design of mobile security notifications and mobile app usability on users' security perceptions and continued use intention. Inf Manag. 2020 Jul;57(5):103235.

Challenges of Trustable AI and Added-Value on Health
B. Séroussi et al. (Eds.)

doi:10.3233/SHTI220620

Key Elements for the Evaluation of mHealth Applications: Results from a Delphi Survey

Felix HOLL[a,b,1] and Walter SWOBODA[a]
[a] *DigiHealth Institute, Neu-Ulm University of Applied Sciences, Neu-Ulm, Germany*
[b] *IBE, Ludwig Maximilian University of Munich, Munich, Germany*

Abstract. Mobile Health (mHealth) applications have seen strong growth in recent years, but they are often not systematically evaluated. A Delphi survey was conducted to identify key elements for the evaluation of mHealth applications. Sixteen experts participated in the study, and the study yielded a list of 79 key elements with expert consensus. Thirty-two elements were in the category of structure quality, 29 in process quality, and 18 in outcome quality. The number of key elements highlights the complexity of conducting systematic evaluations of mHealth applications.

Keywords. Mobile Health, Evaluation, Delphi Survey

1. Introduction

Mobile health (mHealth) applications have experienced strong growth in recent years because of their potential to improve patient care [1]. New policies that introduce the reimbursement for mHealth application through publicly financed healthcare systems, such as in Germany or France, are increasing the availability of mHealth applications [2,3]. However, the number of scientific studies to prove this hypothesis is still too small [4]. There is also a lack of a holistic, standardized, and comprehensive evaluation system, which would increase the validity of studies [5].

Several evaluation tools for medical informatics projects and in general and mobile health applications have been developed and implemented in recent years. Some of these have been published in textbooks, and some have been used in reviews [6–9]. The Mobile App Rating Scale, validated and translated in multiple languages, is a widely used academic evaluation tool for mHealth applications [10–12]. Recommendations for reporting evaluation results are made by STARE-HI [13]. However, despite the breadth of evaluation tools available, most evaluations that are performed are not systematical and only focus on single aspects. As a result, the new evaluation tools for eHealth and mHealth should be developed [14,15].

The goal of the study is to define key elements for the evaluation of mHealth applications through expert consensus as a basis for the development of evaluation tools.

[1] Corresponding Author, Felix Holl, DigiHealth Institute, Neu-Ulm University of Applied Sciences, 89231 Neu-Ulm, Germany; E-mail: felix.holl@hnu.de.

2. Methods

We conducted a web-based Delphi survey from April 08th, 2021, until June 20th, 2021. The survey consisted of three rounds (14 days each) and a break of 2 weeks in-between rounds that were used for data analysis and preparation of the subsequent stages. Participants were recruited through a mailing list of the American Medical Informatics Association and a post in a professional social network. Participation was voluntary, and informed consent for participation and data storage was obtained. The responses of the participants were recorded anonymously. E-mail addresses collected to disseminate the invitations for the subsequent survey rounds were stored in a separate database. Participants had the opportunity to withdraw their consent and stop their participation at any point in the study.

2.1. Eligibility criteria

Subject Matter Experts (SME): were eligible to participate if they met the following requirements:

- Five years or more experience in informatics or medicine or health sciences
- Experience in health IT with evaluation methods or having conducted an evaluation before
- Age 18 years or older, able to read and write English, and access to a computer with internet access.

2.2. Study setup and instruments

The first round of the survey was an open idea generation phase using the leading research question "What are important aspects when evaluating mHealth applications?" The idea generation was followed by two rounds of consensus building. (Visualized in **Figure 1**).

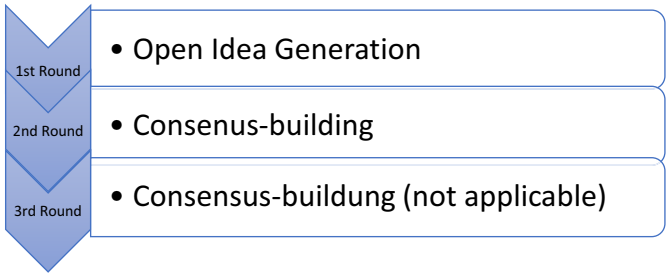


Figure 1. Overview of the study setup

In the open idea generation phase (first round), SMEs were asked to list their ideas on important indicators for the evaluation of mobile health applications that need to be included in a list. The survey consisted of free texts fields, where participants could enter their responses. In addition, demographic information of the participants was collected.

The indicators entered by the SMEs were be grouped based on the Donabedian model of quality (structure, process, and outcome), and duplicates removed before the consensus-building rounds [16]. The Donabedian model of quality was used as a theoretical basis to structure the results of the survey and improve the organization of the

results with the lens of quality of care and categorize elements of app quality according to their influence on quality of care. Following the idea generation phase, two rounds of consensus building took place in which the participants were asked to rate each of the indicators from the idea generation phase using a 4-point Likert scale (Strongly agree, agree, disagree, strongly disagree) whether they should be included in the list of key elements. In addition to the rating, participants could suggest additional elements or rewording of elements. Consensus was considered if at least 75% of the participants rated an element with "agree" or "strongly agree". Elements with less than 75% agreement were dropped. Studies have found that 75% is a common level of agreement in Delphi surveys [17]. Because no additional elements were suggested in the first consensus-building cycle, the study was stopped after the 2nd round. Data cleaning was done by one author and validated by a second author. Data analysis was done using Microsoft Excel Version 16.

3. Results

3.1. Expert demographics

A total of 16 experts participated in the web-based Delphi survey. All participants met the experience criteria for participating in the study. All participants fulfilled the eligibility criteria of at least five years of professional experience. Participants came from several different professional backgrounds.

- User Adoption
- Epidemiology & Public Health
- Public health
- Medical informatics
- Health information management
- Law; Health technology regulation

Three participants had 5 to 7 years professional experience, one 8 to 10, four 11 to 15, four 16 to 20 and another four more than 20 years.

3.2. Results of the open idea generation phase (Round 1)

The idea generation in round 1 yielded a total of 82 elements after the removal of duplicates. Of the 82 elements, 34 were categorized as structure quality, 30 as process quality, and 18 as outcome quality.

3.3. Results of the consensus-building

The first cycle of consensus-building (round 2) resulted in 79 key elements that achieved at least 75% consensus (rated with "agree" or "strongly agree" on the 4-point Likert scale). These 79 elements were divided into 32 elements in the category of structure quality, 29 in the category of process quality, and 18 in the category of outcome quality. Three elements from the open idea generation phase did not achieve consensus. As mentioned, no additional elements were suggested in the second round, and the Delphi survey was stopped after the second round.

Table 1 give an overview of the number of key elements in each quality category and subcategories. The complete list of key elements can be accessed here [18].

Table 1. Structure quality

Structure Quality (n=32)
Data Quality and Interoperability (n=4)
Privacy (n=5)
Security (n=7)
Funding/Cost (n=3)
Access (n=3)
Functionality (n=9)
Certification (n=1)
Process Quality (n=29)
UX (n=9)
Usability (n=3)
Quality improvement (n=4)
Content (n=7)
Usage (n=6)
Outcome Quality (n=18)
Health outcomes (n=9)
Economic outcomes (n=2)
Care process outcomes (n=1)
Patient-reported outcomes (n=5)
Other outcomes (n=1)

4. Discussion

Seventy-nine key elements achieved expert consensus in the Delphi survey. The elements are spread out across all three of the Donabedian model of quality. Structure quality has the most elements while outcome quality has the least, as the outcome quality can be assessed with few indicators. In contrast, the structure quality is determined by several elements. The usual usability studies on mHealth apps [19] should be followed in the future by content studies based on key elements such as those shown here.

5. Conclusions

The high number of elements highlights the complexity of conducting holistic, standardized, and comprehensive evaluations of mHealth applications. The results from the Delphi survey can serve as a basis for the development of a comprehensive evaluation framework for mHealth applications. The next step is to search for validated measurement tools for each key element and create new tools for those where no validated ones exist. Afterward, the framework will be developed and piloted.

References

[1] Steinhubl SR, Muse ED, Topol EJ. The emerging field of mobile health. Sci Transl Med. 2015 Apr 15;7(283):283rv3. doi: 10.1126/scitranslmed.aaa3487. PMID: 25877894; PMCID: PMC4748838.

[2] Gerke S, Stern AD, Minssen T. Germany's digital health reforms in the COVID-19 era: lessons and opportunities for other countries. NPJ Digit Med. 2020 Jul 10;3:94. doi: 10.1038/s41746-020-0306-7. PMID: 32685700; PMCID: PMC7351985.

[3] Yaghobian S, Ohannessian R, Duong TA, Medeiros de Bustos E, Le Douarin YM, Moulin T. France extends its tele-expertise funding model nationally after COVID-19. J Telemed Telecare. 2022 Apr;28(3):233-235. doi: 10.1177/1357633X211067067. Epub 2021 Dec 21. PMID: 34931877.
[4] Rowland SP, Fitzgerald JE, Holme T, Powell J, McGregor A. What is the clinical value of mHealth for patients? NPJ Digit Med. 2020 Jan 13;3:4. doi: 10.1038/s41746-019-0206-x. PMID: 31970289; PMCID: PMC6957674.
[5] Holl F, Kircher J, Swoboda WJ, Schobel J. Methods Used to Evaluate mHealth Applications for Cardiovascular Disease: A Quasi-Systematic Scoping Review. Int J Environ Res Public Health. 2021 Nov 23;18(23):12315. doi: 10.3390/ijerph182312315. PMID: 34886039; PMCID: PMC8656469.
[6] Friedman CP, Wyatt J. Evaluation methods in biomedical informatics. Springer Nature; 2022.
[7] Brender J. Handbook of evaluation methods for health informatics. London: Elsevier Academic Press; 2006.
[8] Eslami Andargoli A, Scheepers H, Rajendran D, Sohal A. Health information systems evaluation frameworks: A systematic review. Int J Med Inform. 2017 Jan;97:195-209. doi: 10.1016/j.ijmedinf.2016.10.008. Epub 2016 Oct 15. PMID: 27919378.
[9] Fanta G, Pretorius L, Erasmus L. An evaluation of ehealth systems implementation frameworks for sustainability in resource constrained environments: A literature review. InInternational Association for Management of Technology (IAMOT) 2015 Conference Proceedings 2015 Jun 8 (pp. 1046-1063).
[10] Stoyanov SR, Hides L, Kavanagh DJ, Zelenko O, Tjondronegoro D, Mani M. Mobile app rating scale: a new tool for assessing the quality of health mobile apps. JMIR Mhealth Uhealth. 2015 Mar 11;3(1):e27. doi: 10.2196/mhealth.3422. PMID: 25760773; PMCID: PMC4376132.
[11] Terhorst Y, Philippi P, Sander LB, Schultchen D, Paganini S, Bardus M, Santo K, Knitza J, Machado GC, Schoeppe S, Bauereiß N, Portenhauser A, Domhardt M, Walter B, Krusche M, Baumeister H, Messner EM. Validation of the Mobile Application Rating Scale (MARS). PLoS One. 2020 Nov 2;15(11):e0241480. doi: 10.1371/journal.pone.0241480. PMID: 33137123; PMCID: PMC7605637.
[12] Knitza J, Tascilar K, Messner EM, Meyer M, Vossen D, Pulla A, Bosch P, Kittler J, Kleyer A, Sewerin P, Mucke J, Haase I, Simon D, Krusche M. German Mobile Apps in Rheumatology: Review and Analysis Using the Mobile Application Rating Scale (MARS). JMIR Mhealth Uhealth. 2019 Aug 5;7(8):e14991. doi: 10.2196/14991. PMID: 31381501; PMCID: PMC6699116.
[13] Talmon J, Ammenwerth E, Brender J, de Keizer N, Nykänen P, Rigby M. STARE-HI--Statement on reporting of evaluation studies in Health Informatics. Int J Med Inform. 2009 Jan;78(1):1-9. doi: 10.1016/j.ijmedinf.2008.09.002. Epub 2008 Oct 18. PMID: 18930696.
[14] Catwell L, Sheikh A. Evaluating eHealth interventions: the need for continuous systemic evaluation. PLoS Med. 2009 Aug;6(8):e1000126. doi: 10.1371/journal.pmed.1000126. Epub 2009 Aug 18. PMID: 19688038; PMCID: PMC2719100.
[15] Enam A, Torres-Bonilla J, Eriksson H. Evidence-Based Evaluation of eHealth Interventions: Systematic Literature Review. J Med Internet Res. 2018 Nov 23;20(11):e10971. doi: 10.2196/10971. PMID: 30470678; PMCID: PMC6286426.
[16] Donabedian A. The quality of care. How can it be assessed? JAMA. 1988 Sep 23-30;260(12):1743-8. doi: 10.1001/jama.260.12.1743. PMID: 3045356.
[17] Foth T, Efstathiou N, Vanderspank-Wright B, Ufholz LA, Dütthorn N, Zimansky M, Humphrey-Murto S. The use of Delphi and Nominal Group Technique in nursing education: A review. Int J Nurs Stud. 2016 Aug;60:112-20. doi: 10.1016/j.ijnurstu.2016.04.015. Epub 2016 May 1. PMID: 27297373.
[18] Holl F. Key Elements for the Evaluation of mHealth Applications, (2022). doi:doi/10.7910/DVN/D5XR0Q.
[19] Reza Soroushmehr SM, Davuluri P, Molaei S, Hargraves RH, Tang Y, Cockrell CH, Ward K, Najarian K. Spleen segmentation and assessment in CT images for traumatic abdominal injuries. Journal of medical systems. 2015 Sep;39(9):1-1.

Challenges of Trustable AI and Added-Value on Health
B. Séroussi et al. (Eds.)

doi:10.3233/SHTI220621

Personalising Symptoms Reporting in Telemonitoring Applications for Cancer Patients

Lucia SACCHI[a], Nicole VEGGIOTTI[a], Emanuele GIRANI[a], Valentina TIBOLLO[b], Silvana QUAGLINI[a,1]
[a] *Dept of Electrical, Computer and Biomedical Engineering, University of Pavia, Italy*
[b] *LISRC laboratory, Istituti Clinici Scientifici Maugeri SpA SB IRCCS, Pavia, Italy*

Abstract. Patient reported outcomes have been shown to be predictive of cancer patients' prognosis, and their monitoring through electronic applications have been shown to positively impact survival. On the other hand, patient apps in general show a number of criticalities that often lead patients to abandon their use. One of them is usability. A scarce attention to usability during app development leads to unsatisfactory user experience. In this work, we present an algorithm to facilitate patient symptoms reporting, by personalising the list of symptoms according to their probability of occurrence in the specific patient. This avoids searching long lists of items, thus decreasing the patients' burden in symptom reporting.

Keywords. Patient reported outcomes, cancer, interface terminology, telemedicine, personalised medicine.

1. Introduction

There is increasing evidence on the benefits of telemonitoring systems that allow cancer home patients to report adverse events during oncological treatments [1][2][3]. Some of them [4] start being approved as "digital therapeutics" by regulatory organisations such as FDA. Using those systems, patients can enter symptoms as soon as they appear, and this represents two advantages. First, reporting is more accurate, because the system may ask patients to enter details that could be forgotten if asked by the physician at the next visit, and second, doctors can see what's happening in-between control visits. In fact, type, severity, and duration of symptoms are essential for both doctors and decision support algorithms, for a correct interpretation and management of adverse drug events (ADE). Thus, telemonitoring apps should offer any possible facility to maximize the patient's compliance with accurate symptom reporting. Moreover, while reporting could be clear and precise even using free text, using structured data is advisable, to allow easier and faster electronic data elaboration. Therefore, first of all, an interface terminology must be chosen, including all the possible symptoms a patient could experience, which is a very high number, and then users must be provided with facilities for quickly searching the symptom(s) to be entered. To make some examples, the symptom list can be shown as a flat list in alphabetical order, or symptoms can be

[1] Corresponding Author, Silvana Quaglini; E-mail: silvana.quaglini@unipv.it.

grouped according to the body district affected or to the physiological system involved, or patients could start writing a text and the *autocomplete* function looks for the possible compatible labels, etc. In any case, even after those filters, the number of remaining symptoms, among which the patient should choose, could be uncomfortably high. Since there is evidence that ADEs frequency depends on cancer type, treatment, and treatment duration, in this paper we present an algorithm for sorting those remaining symptoms according to their probability of occurrence, for *that* patient at *that* time, in order to maximize the chance for a patient to find the symptom among the top ones in the list. This will improve the user experience with the app, thus increasing the chance of using it over a long period. The paper describes also the data model the algorithm runs on top of, which stores information collected from both the patient's profile and from the literature. Since this work is part of a European project, the next section will briefly illustrate the project objectives.

2. The CAPABLE project

CAPABLE is a EU Horizon2020 project (Jan 2020-Dec 2023), currently in the middle of its development, which implements an overall intervention strategy for improving cancer patients' wellbeing, both physical and mental. It helps increase patients' awareness about their condition, understand and cope with daily needs, become more proactive and more positive in their cancer journey. To this purpose, patients will be provided with a smartwatch (for the automatic acquisition of physical activity and some vital parameters), and an app for reporting symptoms and answering some follow-up questionnaires. The app will also send recommendations to patients, based on scientific evidence and/or approved by the CAPABLE experts panel, and suggest some exercises, both physical and mental, to achieve objectives that patients, at the enrollment, may have set together with their oncologist. CAPABLE targets also the multidisciplinary healthcare team who takes care of patients, namely oncologists, psychologists and nutritionists. Doctors will rely on a web interface that will visualise their home patients' data and will suggest evidence-based interventions for preventing and managing adverse events. Since this paper deals with symptoms reporting, the next section will describe this functionality in more detail.

For structured symptom reporting, the project relies on the Common Terminology Criteria for Adverse Events (CTCAE), developed by NIH, nowadays used at an international level to represent ADEs. More precisely, we use a subset of 130 terms, obtained by excluding those events that cannot be noticed by patients or their caregivers (for example toxicities that can be detected only by diagnostic laboratory tests) or that are not of interest for cancer patients, according to medical experts' opinion. To further filter the symptoms at runtime, a body-shaped graphical interface allows indicating the body part affected, for example choosing the head will filter out all symptoms related only to limbs or torax. Finally, the autocomplete function is available. These functionalities, present also in other applications [5], are useful to shorten the symptom list that a patient has to examine on his small smartphone interface, for selecting the specific symptom he wants to enter. However, they do not take into account that symptom incidence varies according to cancer, treatments and time. In the following, we show how we exploited literature data about ADE incidence to build an algorithm that considers also those aspects, and that delivers a final, dynamic and patient-specific symptom list, which is more likely to visualize the most probable ADEs at the top.

3. Evidence about Adverse Events of Oncological Treatments

The literature offers several data about the overall incidence of ADEs that occur for specific cancer patients undergoing specific treatments. For our proof of concept, we limited our search to the treatments currently used according to the ESMO clinical guidelines (www.esmo.org/guidelines) for melanoma and renal cell cancer, which are the two main pathologies considered in CAPABLE. We relied on phase III clinical trials, being meant (also) to assess the risk profile of the drugs on large samples, thus putting particular attention to the ADE occurrence [6][7][8][9][10]. For several ADE types, in addition to overall incidence, we found interesting information about time courses [6][11][12]. As an example, Figure 1 shows the time-related onset probability of some ADEs caused by immunotherapy with nivolumab. It can be noticed that: (a) ADEs appear in different times; (b) the first weeks are the most affected; (c) some ADEs are more frequent in the early treatment phase but can also occur very later.

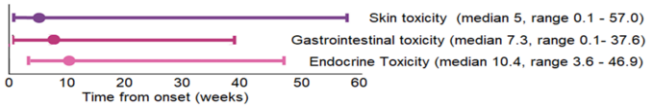


Figure 1. Onset time (median and range) for different ADEs caused by nivolumab - Adapted from [6]

Similar information can be found for ADEs related to other active principles. In addition to incidence, it's important to estimate the event duration: as a matter of fact, recovery time may last from a few days to months [6].

4. Methods

We used literature derived data such as the ones shown in Figure 1, to simulate patterns of time variant incidence for each ADE considered. We have chosen the lognormal distribution, since it is defined on a positive values domain, it may account for outliers (in our case very late ADEs), and its two parameters μ and σ allow us to shape the peak position and kurtosis appropriately. Thus, this distribution family is suitable to represent asymmetrical distributions, with the peak on the left and different time spans. Figure 2 (left) shows the Matlab code allowing to assess a specific distribution (namely the skin toxicity), while in the right the simulated distributions are shown for the considered ADEs, using *week* as a suitable time granularity.

```
%% skin

% interval on which the pdf of the lognormal distribution is evaluated
x = 0:0.1:60;
t0 = 1;
t1 = 60;

% mu coefficient, equal to the natural logarithm of the median
mu_skin = log(6.1);

% sigma coefficient
sigma_skin = 0.8;

% pdf evaluation in the interval x
y_skin = lognpdf(x,mu_skin,sigma_skin);

% pdf plot
figure('name','pdf - skin')
plot(x,y_skin)

% pdf explicit expression
f = @(y) 1./(sigma_skin.*y*sqrt(2*pi)) ...
    .* exp(-(log(y)-mu_skin).^2/(2*sigma_skin^2));
% weekly coefficients vector
w_skin = zeros(60,1);

% weekly coefficients evaluation as the probability that ADEs
% of the skin category occur during week t
for t = t0:t1
    w_skin(t) = integral(f,t-1,t);
end
```

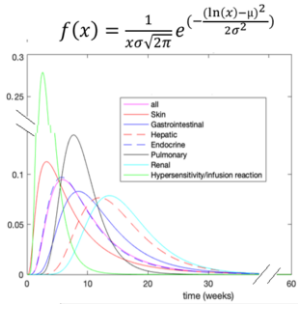


Figure 2. (Left) Matlab code used to approximate onset time distribution of a specific ADE category, and (right) the lognormal distribution formula and the obtained distribution models

In particular, the *for* cycle in the bottom of the Matlab code fills in the array w_skin with the integral of the probability density calculated within each week from the therapy start. Those values are a sort of "weekly coefficients", representing the probability that an ADE occurs within each week (for sake of demonstration we limited the time horizon to 60 weeks), under the hypothesis that total probability is 1. However, since each ADE probability is less than one, those coefficients are multiplied by the actual overall probability of occurrence found in the literature as described in the above section.
Both data gathered from the literature and the calculated weekly coefficients have been structured into a relational database (DB), illustrated in Figure 2 through its entity-relationship (ER) diagram. The publications that have been consulted to retrieve data from clinical studies are listed in the *Paper* entity through their DOI. *Cohort* represents the cohorts that have been studied in the selected papers. Each cohort is defined by a *CohortType*, which is a combination of type of cancer and therapy. In some rows the therapy does not correspond to a specific treatment (e.g. Nivolumab monotherapy, Nivolumab plus Ipilimumab) but is set as "pooled". In this case the CohortType is a dummy, since its only purpose is to gather from the database the weekly coefficients for a pool of treatments when the coefficients for a particular cancer-therapy pairing are not available; this procedure only works when the treatment is included in the *TreatmentPool* for that cohort type (i.e., if the DB contains weekly coefficients for a pool of treatments including drug A and the combination of drugs A and B, those coefficients will not be used for drug C). *AdverseEvent* contains the list of ADEs, whose frequency of occurrence in each cohort is stored in *Frequency*. The entity *Category* contains the categories to which the adverse events belong. This is important to link adverse events to weekly coefficients, that have been calculated for ADEs categories, and are contained in *Coefficient*.

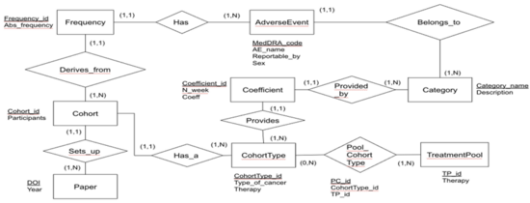


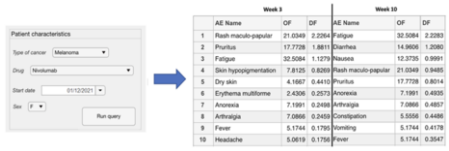
Figure 3. ERdiagram of the database used for storing literature data about ADE incidence

This DB is exploited by the symptom prioritization query described in the next section.

5. Results

To demonstrate the potentiality of the algorithm, we developed a lite Matlab interface, which allows entering the variables that affect ADE incidence, including the treatment type and its start date (Figure 3 left). A set of SQL queries is then run on the DB, and both overall and dynamic frequency (OF and DF) of the symptoms are generated. Figure 3 (right) shows the results of the algorithm when run for weeks 3 and 10 from Start Date. The OF values account for the incidence over the full treatment period, while DF also for the time distribution, thus referring to the single actual week considered. In the figure, we sorted the symptom list according to DF. The ordering that would have been generated considering OF is remarkably different. In our example, a melanoma patient treated with Nivolumab, Fatigue is overall the most frequent ADE, but if we take into

account time distribution, this is true only after certain time (week 10 in our example), while in week 3 other symptoms like maculo-papular rash are more likely to arise.



	Week 3			Week 10		
	AE Name	OF	DF	AE Name	OF	DF
1	Rash maculo-papular	21.0349	2.2364	Fatigue	32.5084	2.2283
2	Pruritus	17.7728	1.8811	Diarrhea	14.9606	1.2080
3	Fatigue	32.5084	1.1279	Nausea	12.3735	0.9991
4	Skin hypopigmentation	7.8125	0.8269	Rash maculo-papular	21.0349	0.9485
5	Dry skin	4.1667	0.4410	Pruritus	17.7728	0.8014
6	Erythema multiforme	2.4306	0.2573	Anorexia	7.1991	0.4935
7	Anorexia	7.1991	0.2498	Arthralgia	7.0866	0.4857
8	Arthralgia	7.0866	0.2459	Constipation	5.5556	0.4486
9	Fever	5.1744	0.1795	Vomiting	5.1744	0.4178
10	Headache	5.0619	0.1756	Fever	5.1744	0.3547

Figure 4. The different ordering of ADEs when considering the time-variant incidence. OF = Overall Frequency; DF = Dynamic Frequency

6. Discussion and Conclusions

This paper is a proof of concept (PoC) of how personalized medicine can be implemented in a digital therapeutics. We proposed an interface terminology that varies over time according to the probability of ADE occurrence, in order to improve the user experience with the app. As a PoC, our study has some limitations. The whole ADE list accounts for 130 items, but we did not collect time-variant data for all of them. On the other hand, the papers considered in this work deal with the most frequent ADEs, so we think that results can provide a good idea of the real-world effectiveness of our algorithm.

Acknowledgments
This work has been funded by the EU Horizon 2020 project CAPABLE, #875052

References

[1] Denis F, et al. Two-Year Survival Comparing Web-Based Symptom Monitoring vs Routine Surveillance Following Treatment for Lung Cancer. JAMA. 2019 Jan 22;321(3):306-307.
[2] Mody GN, et al. Electronic patient-reported outcomes monitoring during lung cancer chemotherapy: A nested cohort within the PRO-TECT pragmatic trial (AFT-39). Lung Cancer. 2021 Dec;162:1-8.
[3] Basch E, et al. Clinical Utility and User Perceptions of a Digital System for Electronic Patient-Reported Symptom Monitoring During Routine Cancer Care: Findings From the PRO-TECT Trial. JCO Clin Cancer Inform. 2020 Oct;4:947-957.
[4] Oleena, (www.oleena.com)
[5] Zini EM, et al. Standardization of immunotherapy adverse events in patient information leaflets and development of an interface terminology for outpatients' monitoring. J Biomed Inform. 2018;77:133-44.
[6] Weber JS, et al. Safety Profile of Nivolumab Monotherapy: A Pooled Analysis of Patients With Advanced Melanoma. J Clin Oncol. 2017 Mar;35(7):785-792.
[7] Larkin J, et al. Combined Nivolumab and Ipilimumab or Monotherapy in Untreated Melanoma. N Engl J Med. 2015 Jul 2;373(1):23-34.
[8] Rini BI, et al. KEYNOTE-426 Investigators. Pembrolizumab plus Axitinib versus Sunitinib for Advanced Renal-Cell Carcinoma. N Engl J Med. 2019 Mar 21;380(12):1116-1127.
[9] Motzer RJ, et al. Avelumab plus Axitinib versus Sunitinib for Advanced Renal-Cell Carcinoma. N Engl J Med. 2019 Mar 21;380(12):1103-1115.
[10] Motzer RJ, et al. CheckMate 214 Investigators. Nivolumab plus Ipilimumab versus Sunitinib in Advanced Renal-Cell Carcinoma. N Engl J Med. 2018 Apr 5;378(14):1277-1290.
[11] Tang SQ, et al. The Pattern of Time to Onset and Resolution of Immune-Related Adverse Events Caused by Immune Checkpoint Inhibitors in Cancer: A Pooled Analysis of 23 Clinical Trials and 8,436 Patients. Cancer Res Treat. 2021 Apr;53(2):339-354.
[12] Martins F, et al. Adverse effects of immune-checkpoint inhibitors: epidemiology, management and surveillance. Nat Rev Clin Oncol. 2019 Sep;16(9):563-580.

Challenges of Trustable AI and Added-Value on Health
B. Séroussi et al. (Eds.)

doi:10.3233/SHTI220622

Changes in Patient Characteristics in Telehealth Usage During Times of Crisis

Saif KHAIRAT[a,1], Ashlyn ZEBROWSKI[a], Elizabeth KWONG[a], Barbara EDSON[b] and Robert GIANFORCARO[b]
[a] *University of North Carolina at Chapel Hill, USA*
[b] *UNC Health, USA*

Abstract. There is limited knowledge on whether increased telehealth usage may enhance health access to communities during natural disasters, particularly for emergency medical services. This study aimed to elucidate telehealth usage during three hurricanes in NC between 2018 and 2020 and assessed demographics of users including gender and age, insurance status, and daily rate of visits in relation to respective hurricanes. From 10,056 telehealth visits, we found that age and insurance coverage were significantly different between crisis and non-crisis times. Patients found comparative satisfaction during both times. This study suggests the use of phone and video visits to enable better access to parents with children under the age of 18 years and uninsured patients.

Keywords. Patient, Telehealth, Crisis

1. Introduction

Hurricanes are natural disasters that take a large toll on individuals and infrastructure, causing damage to property, loss of power and resources, and potentially death. These disasters have short- and long-term impacts on the access to and quality of healthcare and individual health, including morbidity and mortality, disease burden, disruption in healthcare delivery, and mental health. In addition to limiting and delaying patients' access to healthcare, hurricanes impact healthcare communication, transportation, and other facility disruptions [1]. Between 2005 and 2020, a total of 33 hurricanes hit the United States with a total mortality rate of 2003 individuals, resulting in a mortality rate between 0 and 1,518 per year related to hurricanes [2]. With several hurricane crises impacting U.S. healthcare systems every year, patients need telehealth solutions during times of crisis. In 2018, Hurricane Florence left over 890,000 customers lost power across 76 counties in NC [3]. Florence resulted in a total of 40 deaths in North Carolina [4]. In 2019, over 288,000 customers lost power across eastern North Carolina following Hurricane Dorian [5]. In 2020, about 362,000 power outages were reported in North Carolina, with approximately 45,000 of customers still without power because of Hurricane Isaias [6]. Telehealth provides benefits to patients that include potentially shortened wait times, elimination of travel time, and limited physical contact and

[1] Corresponding Author, Saif Khairat, PhD, MPH, 428 Carrington Hall, Chapel Hill, NC 27514 USA; E-mail: saif@unc.edu.

exposure, which may be beneficial for individuals with health needs during a natural disaster [7].

Telehealth usage has been shown to vary by patient population and demographic [8], and continues to be increasingly utilized and studied. There is also identified potential for telemedicine programs particularly in high-risk areas for disasters [9]. However, research on the impact and usage of telehealth in the US during periods of natural disasters or crises is limited and warrants further studies. To that extent, there is limited knowledge on whether increased telehealth usage may enhance health access to communities during natural disasters, particularly for emergency medical services [10]. This study aimed to elucidate telehealth usage during three hurricanes in NC between 2018 and 2020 and assess demographics of users including gender and age, insurance status, and daily rate of visits in relation to respective hurricanes.

2. Methods

In this cross-sectional study, we examined the usage of telehealth services at a virtual care center (VCC) during three hurricanes: Florence, Dorian, and Isaias. Between August 2018 and August 2020, the three hurricanes struck North Carolina (NC): Hurricane Florence (August 31-September 18, 2018), Hurricane Dorian (August 24-September 10, 2019), and Hurricane Isaias (July 28-August 5, 2020). Hurricanes ranged in number of days, category, amount of rainfall, wind speeds, and pressure.

VCC provides on-demand virtual services for primary and urgent care needs to individuals residing in North Carolina, USA since 2018. Patients have to create an account through a web portal in order to schedule an appointment with a provider. VCC providers were licensed to treat or consult on a wide range of medical conditions and common ailments including ear infections, fevers, rashes, and respiratory infections. Although health insurance is accepted, patients have an option to pay an out-of-pocket flat fee. Patients could schedule the appointment for a later date if they preferred to. This study was approved by the institutional review board.

We extracted visit data from 2018 to 2020, which included patient demographics and date and time of the visit. We categorized each visit to one of four categories: Florence, Dorian, Isaias, and non-Crisis. non-Crisis were the telehealth visits that occurred during time were there were no crisis i.e., hurricanes. We ran descriptive analysis on the data, and we tested if there were any differences in characteristics among patients in times of crisis versus regular times.

3. Results

Of 10,056 telehealth visits, 540 (5.4%) occurred during times of crisis. During crisis times, 383 (70.9%) telehealth visits were from females, 203 (37.6%) were from individuals between the ages of 18-35 years, 390 (72.2%) were from uninsured patients, 283 (52.4%) were from urban regions within North Carolina. On the contrary, during non-Crisis times, there were 7109 (74.6%) visits from females, 3850 (40.4%) were from individuals 18-35 years, 5926 (62.1%) were uninsured patients, and 5339 (56%) were from urban regions, table 1.

Most telehealth visits in NC during each hurricane were from rural areas, with the exception of Hurricane Isaias. During Hurricane Florence, 118/232 (50.9%) patients

were from rural areas. During Hurricane Dorian, 101/154 (65.6%) patients were from rural areas. During Hurricane Isaias, only 64/154 (41.6%) patients were from rural areas. Of the 232 total visits during Hurricane Florence, 188 (81.0%) were uninsured. Of the 154 total visits during Hurricane Dorian, 88 (57.1%) were uninsured. Of the 154 total visits during Hurricane Isaias, 114 (74.0%) were uninsured.

Table 1. Patient Characteristics for all visits, non-crisis visits, crisis visits, and by each hurricane.

	All	Non-Crisis	Crisis	Hurricane Florence	Hurricane Dorian	Hurricane Isaias
Total	10075 (100%)	9535 (94.6%)	540 (5.4%)	232 (2.3%)	154 (1.5%)	154 (1.5%)
Gender						
Female	7492 (74.4%)	7109 (74.6%)	383 (70.9%)	159 (68.5%)	121 (78.6%)	103 (66.9%)
Male	2564 (25.4%)	2407 (25.2%)	157 (29.1%)	73 (31.5%)	33 (21.4%)	51 (33.1%)
Age						
2-17	983 (9.8%)	910 (9.5%)	73 (13.5%)	41 (17.7%)	15 (9.7%)	17 (11.0%)
18-35	4053 (40.2%)	3850 (40.4%)	203 (37.6%)	82 (35.3%)	64 (41.6%)	57 (37.0%)
36-50	3413 (33.9%)	3224 (33.8%)	189 (35.0%)	77 (33.2%)	54 (35.1%)	58 (37.7%)
51-65	1418 (14.1%)	1355 (14.2%)	63 (11.7%)	28 (12.1%)	19 (12.3%)	16 (10.4%)
65+	208 (2.1%)	196 (2.1%)	12 (2.2%)	4 (1.7%)	2 (1.3%)	6 (3.9%)
Insurance Status						
Insured	3759 (37.3%)	3609 (37.9%)	150 (27.8%)	44 (19.0%)	66 (42.9%)	40 (26.0%)
Uninsured	6316 (62.7%)	5926 (62.1%)	390 (72.2%)	188 (81.0%)	88 (57.1%)	114 (74.0%)
Location						
Rural	5622 (56.5%)	5339 (56.0%)	283 (52.4%)	118 (50.9%)	101 (65.6%)	64 (41.6%)
Urban	4334 (43.5%)	4085 (42.8%)	249 (46.1%)	112 (48.3%)	53 (34.4%)	84 (54.5%)

We found significant differences in patient characteristics in telehealth use between times of crisis and non-crisis, table 2. Patients' age (p-value<0.05) and insurance type (p-value < 0.05) were statistically significant during crisis times such that more uninsured patients utilized telehealth during crisis times compared to regular times. Also, there was a significant increase in patients between 2-17 years during times of crisis. Patients' gender (p-value=0.07) and rurality (p-value=0.25) were not significantly different between both times.

Table 2. Chi-Square of the differences between Crisis and non-Crisis times.

Patient Characteristics	Chi-square Value	P-value
Gender	5.1697	0.0754
Age	14.8837	0.0049

Insurance	21.5645	0.000003
Rurality	2.7662	0.2508

Overall, the average daily rate of telehealth visits at VCC was 11.4 visits. During crisis times, the average was 13.3 compared to 11.4 visits during non-Crisis times, figure 1. The highest average daily rate of telehealth visits occurred during Hurricane Isaias (18.1 visits) while the lowest was during Hurricane Dorian (8.8 visits). When comparing the daily rate before COVID-19 and after, the daily number of telehealth visits pre-COVID were 9.8 compared to 20.8 after COVID.

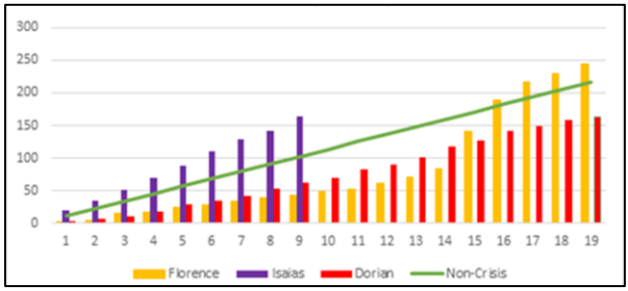


Figure 1. The number of telehealth visits by hurricane (Florence, Dorian, and Isaias) by day from start date (Day 1) to end date (varies) compared to number of telehealth visits during non-crisis times.

There were no significant differences found in patient satisfaction between crisis and non-crisis times. During crisis, 82% of patients who responded to the survey reported an overall positive experience using telehealth compared to 83.6% during non-crisis times. Similarly, the proportion of patients who indicated a negative experience with telehealth were similar between crisis (18%) and non-Crisis (16%), Figure 2.

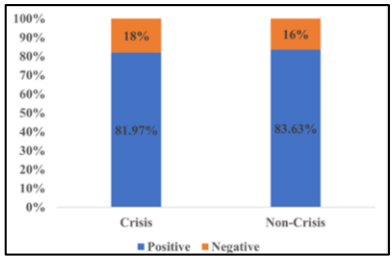


Figure 2. Patient-reported overall telehealth experience stratified by crisis and non-crisis times.

4. Discussion

In this cross-sectional study, we examined the variations in patient characteristics and satisfaction between times of crisis vs. non-crisis. We report that patient age and insurance coverage changed significantly during times of crisis. There was higher utilization of telehealth for patients under 18 years, which may be explained by an increased need for pediatric care when roads and provider offices may be closed during hurricanes. Similarly, the substantial increase in uninsured patients seeking care during crisis was profound. It is plausible that patients with no health insurance may reside in areas with poor infrastructure or no access to care and hence, during crisis when the need

for health care increases, uninsured patients used telehealth to overcome their financial and geographic barriers. The gap in knowledge of telehealth utilization during times of crisis is of importance to policy makers and health organizations as they continue to tailor their telehealth usage. Our findings show that patients in general found merit in using telehealth and particularly, younger patients and uninsured patients. For pediatric and adolescent patients, the ability to seek care without the need to go out during crisis increased their utilization of telehealth. The ability for patients to pay a relatively small out-of-pocket fee for a telehealth consult may have enabled uninsured individuals to seek care during times in crisis. Moreover, we think that providing patients with the option to use phone or video to connect with their provider increased the utilization of specific patient groups such as over 65 years, uninsured, and rural patients. Therefore, we recommend that organizations consider providing both modalities to increase the likelihood of patients who are not able to or are not comfortable enough to use video calls to seek care through telephone. Prior literature has shown that telehealth allows patients and caregivers to overcome physical barriers to convenient medical care [11], and provides a timely solution to provide patient care and services during times of crisis, made particularly evident during the COVID-19 pandemic [12]. However, disparities in telehealth familiarity, knowledge, and willingness to use telehealth still exist, suggesting that further efforts to enhance the use of telehealth effectively and increase access to telehealth are warranted [8]. This study had limitations. We examined a single state-wide telehealth center that belongs to an academic center. Survey response rate was low, which may affect the generalizability of the survey findings. We did not account for the changes in telehealth usage during the COVID-19 pandemic because at the time of this study the pandemic was still ongoing and analyzing its data would not be complete.

References

[1] Waddell SL, et.al. Perspectives on the Health Effects of Hurricanes: A Review and Challenges. International Journal of Environmental Research and Public Health. 2021;18(5):2756.
[2] Insurance Information Institute. Facts + statistics: Hurricanes 2021 [Available from: https://www.iii.org/fact-statistic/facts-statistics-hurricanes.
[3] Rice D. Hurricane Florence power outages top 890,000, could hit 3 million as storm unleashes fury USA Today2018 [Available from: https://www.usatoday.com/story/news/nation/2018/09/14/hurricane-florence-power-outages/1301060002/.
[4] Stewart SR, Berg R. Hurricane Florence (AL062018) 2019 [Available from: https://www.nhc.woc.noaa.gov/data/tcr/AL062018_Florence.pdf.
[5] Duke Energy. Duke Energy restores power to more than 288,000 customers in eastern North Carolina following Hurricane Dorian 2019 [Available from: https://news.duke-energy.com/releases/duke-energy-restores-power-to-more-than-288-000-customers-in-eastern-north-carolina-following-hurricane-dorian.
[6] Latto A, Hagen A, Berg R. Hurricane Isaias (AL092020) 2021 [Available from: https://origin.www.nhc.noaa.gov/data/tcr/AL092020_Isaias.pdf.
[7] U.S. Department of Health & Human Services. What is telehealth? 2021 [Available from: https://telehealth.hhs.gov/patients/understanding-telehealth/.
[8] Fischer SH, Ray KN, Mehrotra A, Bloom EL, Uscher-Pines L. Prevalence and Characteristics of Telehealth Utilization in the United States. JAMA network open. 2020;3(10):e2022302-e.
[9] Crutchfield AS, Harkey KA. A comparison of call volumes before, during, and after Hurricane Harvey. The American journal of emergency medicine. 2019;37(10):1904-6.
[10] Latifi R, Tilley EH. Telemedicine for disaster management: can it transform chaos into an organized, structured care from the distance? Am J Disaster Med. 2014;9(1):25-37.
[11] Wosik J, Fudim M, Cameron B, Gellad ZF, Cho A, Phinney D, et al. Telehealth transformation: COVID-19 and the rise of virtual care. J Am Med Inform Assoc. 2020;27(6):957-62.
[12] Paterson C, et al. The Role of Telehealth During the COVID-19 Pandemic Across the Interdisciplinary Cancer Team: Implications for Practice. Semin Oncol Nurs. 2020;36(6):151090.

Challenges of Trustable AI and Added-Value on Health
B. Séroussi et al. (Eds.)

doi:10.3233/SHTI220623

Evaluation of the Efficiency of Telemedicine in the Management of Cardiovascular Diseases in Primary Healthcare in Sub-Saharan Africa: A Medico-Economic Study in Cameroon

Georges BEDIANG[a,1], Chris Nadège NGANOU-GNINDJIO[a], Yannick KAMGA[b], Jean Serge NDONGO[a], Fred-Cyrille GOETHE DOUALLA[a], Cheick Oumar BAGAYOKO[c] and Samuel NKO'O[a]

[a] *Faculty of Medicine and Biomedical Sciences, University of Yaoundé I, Cameroon*
[b] *Department for the Control of Disease, Epidemics and Pandemics, Ministry of Public Health, Cameroon*
[c] *Faculté de Médecine, de Pharmacie et d'Odonto-Stomatologie, University of Bamako, Mali*

Abstract. Objective: To assess the efficiency of tele-expertise (tele-ECG) for patients and for health facilities in managing patients with cardiovascular diseases (CVDs) in primary health care in Cameroon. Method: This study was a medico-economic study combining two approaches: cost minimization and cost-effectiveness analysis. It was conducted alongside the previous published controlled multicenter study carried out in Cameroon's two health facilities where tele-ECG has been implemented (intervention centres) and two other where telemedicine has been not implemented (control centres). Results: The average total cost for patients was 9 286 F CFA (US$: 16) in the intervention centres compared to 28 357 F CFA (US$: 49) in the control centres ($p < 0.01$). The calculated ICER favouring tele-ECG was 25 459.6 F CFA (US$: 44). Discussion: Telemedicine is efficient for managing patients with CVDs in primary health care in Cameroon. It enables health facilities in remote areas to offer new healthcare services at a lower cost and improve patients' financial access to healthcare.

Keywords. Telemedicine, Efficiency, Cardiovascular Diseases, Cameroon, Sub-saharan Africa, Developing Countries, Low-middle Income Countries.

1. Introduction

Cardiovascular diseases (CVDs) are one of the major causes of death in the world [1] and represent a significant economic burden on health care systems [2]. In Cameroon, their mortality rate is about 12% [3]. This high mortality rate can be explained by the difficult access (geographical and financial) to care and the lack of healthcare professionals, infrastructures and governance [4]. With the rapid development of digital health, telemedicine can strengthen the health system by offering new possibilities for

[1] Corresponding Author, Georges BEDIANG, PO Box: 1364, Faculty of Medicine and Biomedical Sciences, University of Yaoundé I, Yaoundé – Cameroon; E-mail: bediang@yahoo.com

the management of patients with CVDs and the development of equitable access to healthcare, particularly in remote areas of developing countries. In addition, several studies have demonstrated its efficiency (ability to reduce patient care costs, good cost-effectiveness ratio) in other setings [5-8]. Following the first study we carried out and related to the effectiveness of tele-ECG [9], this study aimed to evaluate the efficiency of tele-expertise (tele-ECG) for patients and health facilities in the management of patients with CVDs in primary health care in Cameroon, sub-Saharan Africa.

2. Methodology

2.1. Study design

This is a medico-economic study combining cost minimization analysis (for evaluation of efficiency in patients) and cost-effectiveness analysis (for evaluation of efficiency in health facilities). This study was conducted alongside the previous published controlled multicenter study carried out in 2016 and 2017 in Cameroon's four health facilities: two intervention centres (Mbouda and Akonolinga District Hospitals) where tele-ECG was implemented and two control centres (Foumbot and Sa'a District Hospitals) where it was not implemented. Details of the methodology used in this study can be found in the previous publication [9].

2.2. Intervention and control centres

Intervention centres
Patients enrolled received usual primary healthcare: a conventional clinical exam carried out by a local healthcare provider, requests for diagnostic tests, management plan of healthcare and scheduling of appointments. Additionally, patients proposed to perform a resting ECG on-site with the Universal ECG^{TM} (a portable 12-channel ECG device associated with dedicated software enabling them to perform ECG tests, archive them, and print the report in PDF format). The ECG test was systematically associated with a local healthcare provider's request to a cardiologist (expert) for remote expertise via an internet tele-expertise platform Bogou [10]. The local healthcare provider had to send patients' data (demographic and clinical data, diagnostic hypothesis, ECG tests and current management plan) via this platform. Based on this data analysis, the cardiologist provided adapted expertise to better manage patients within 72 hours maximum.
Control centres
In the control centres, patients received the usual primary care only without tele-ECG. For all patients, we recommended a consultation at a cardiology centre or to meet a cardiologist in the nearest town: Bafoussam for Foumbot and Yaoundé for Sa'a.

2.3.Assessment of the effectiveness

The primary outcome (effectiveness) was to evaluate the rate of patients' access to an ECG test and to a cardiologist' expertise (number of patients who performed an ECG test and received a cardiologist's expertise divided by the number of recruited patients) at three months of follow-up.

2.4.Medico-economic assessment

Cost minimization analysis
Patients from different centres were subjected to spending diaries. They were asked to gradually record their spending and further data during the care process to identify various direct (directly linked to the intervention) and indirect (financial consequences of the intervention) costs. In intervention centres, the direct costs were: costs of the local consultation, ECG test and transport (from home to the local hospital). The indirect costs were: the waiting time in hours (time between arrival at the hospital, consultation and completion of the ECG test; 1h was estimated as equivalent to a loss of 1 000 F CFA - US$: 1.7), the cost of communication credit consumption and other expenses related to incidents. In the control centres, the direct costs were: local consultation, ECG test, specialist's consultation, transport (from home to the local hospital and from home to referral hospital), accommodation and feeding. The indirect costs were: the waiting time (at the local healthcare provider and at the specialist), the cost of communication credit consumption and other expenses related to incidents.

Cost-effectiveness analysis
Each hospital was subjected to an economic questionnaire to assess the expenses related to the tele-ECG intervention. It aimed to identify fixed (which do not vary with volume of activities) and variable (which vary according to the volume of activities) costs. In intervention centres, the fixed costs were related to tele-ECG. They included: fees, accommodation, nutrition and transport of staff trainers on the use of the ECG and the Bogou platform and the cost of material (ECG, computer, printer, internet modem). Variable costs included: the cost of electrical energy consumption (related to the production and sending of the ECG and to the interpretation and printing of the ECG), the cost of consumables (ink, ballpoint pen, paper, gel, toilet paper), the cost of coordination of tele-ECG activities, the cost of equipment maintenance, the cost of communication credit consumption (calls and SMSs between local healthcare providers, experts and tele-ECG activities coordinator) as well as the cost of their internet connection. As control centres did not use tele-ECG, their costs were therefore, considered zero. The efficiency of health facilities was calculated based on the Incremental Cost-Effectiveness Ratio (ICER) according to the formula [ICER = (C_{Int} - C_{Con}) / (E_{Int} - E_{Con})]. C_{Int} and C_{Con}: the average of total costs of intervention and control centres; E_{Int} and E_{Con}: effectiveness in intervention and control centres.

2.5. Data analysis and ethics

Appropriate statistical tests and relative risk (RR) were performed for the comparison between the two groups. The study received an ethical clearance CE00398/N°CRERSHC/2016 issued by the Centre Regional Ethics Committee for Human Health Research.

3. Results

About the Sociodemographic and clinical characteristics of participants, one hundred seventy-one participants were recruited, 93 (54.4%) in the intervention centres and 78 (45.6%) in the control centres. In the intervention centres, 57% of participants were

women, and the mean age was 59.3±12.3 years. The main cardiovascular risk factors found were: age (84.9%) and High Blood Pressure -HBP- (48.4%), whereas evaluation of HBP (24.7%) and exertional dyspnea (18.3%) were the main complaints. In the control centres, 55.1% of participants were women, and the mean age was 62.4 ± 12.2 years. The main cardiovascular risk factors found were: age (73.1%) and HBP (56.4%), while evaluation of exertional dyspnea (37.2%) and HBP (14.1%) were the main complaints. There was no statistically significant difference between the two groups except for the primary complaint related to exertional dyspnea. Regarding the Effectiveness (primary outcome), ninety-two (98.9%) of participants had access to an ECG test associated with expertise in the intervention centres. In comparison, 26 (33.3%) only had it in the control centres with a p <0.01 (table 1) and a relative risk equal to 2.97 [CI 95%: 2.17 - 4.06].

Table 1. Effectiveness (primary outcome)

Primary outcome	Intervention centres (N=93)		Control centres (N=78)		p
	n	%	N	%	
ECG + Expertise	92	98.9	26	33.3	<0.01

Regarding patient efficiency, the average total cost for patients was 9 286 F CFA (US$: 16) in the intervention centres compared to 28 357 F CFA (US$: 49) in the control centres (p <0.01), see table 2. The actual costs for the implementation of tele-ECG in intervention centres are shown in table 3. The average total cost in these intervention centres was 1 680 336 F CFA (US$: 2900). As control centers do not use tele-ECG, their costs were considered zero. Considering the parameter of effectiveness, the calculated ICER was 25 459.6 F CFA (US$: 44).

Table 2. Comparison of costs by participants in different centres

Costs	Intervention centres (N=92)		Control centres (N=26)		p
	Mean±SD	Min-Max	Mean±SD	Min-Max	
Direct costs (F CFA)	6575±566	5600-6900	18350±3272	8600-23100	<0.01
Indirect costs (F CFA)	2711±1122	1000-6000	10187±3979	4000-20000	<0.01
Total costs (F CFA)	9286±1504	6600-11600	28357± 5918	16600-43100	<0.01

4. Discussion

For telemedicine to become truly anchored as a credible alternative to usual care [11], evidence of its added value in terms of effectiveness and efficiency is essential. After carrying out the study that demonstrated telemedicine's effectiveness in managing patients with CVDs in Cameroon [5], this study shows its efficiency in promoting financial access to healthcare for patients and improving its efficiency for the provision of care in health facilities. This study shows that participants spent an average of 9 286 F CFA to obtain an ECG and expertise in intervention centres, compared to 28 357 F CFA in control centres, for a savings of 19 071 F CFA (US$: 33 or 67.3%). These results are similar to those of Bagayoko et al. in Mali, who assessed the medico-economic contribution of telemedicine (tele-ultrasound in obstetrics and tele-ECG in cardiology) in remote areas and found that patients who used telemedicine spent less than those who

used the standard practice (19 000 F CFA (US$: 33) against 54 000 F CFA (US$: 93)) [5]. The standard practice was the displacement of these populations towards the capital to benefit from specialized care. Van Os-Medendorp et al. in the Netherlands had also found that in the intervention centres, patients spent less in direct and indirect costs (US$: 3 378) than those in the control centres (US$: 3 972) to benefit from tele-expertise for the treatment of atopic dermatitis [6]. About health facilities, we observed that with an investment of only 25 459.6 F CFA per patient in telecardiology, the State or a health facility would guarantee its population the possibility of performing an ECG and having the expertise of a remote cardiologist in a primary care setting. Like other studies [7, 8], this result shows that telemedicine is cost-effective, especially in remote areas, if we compare it to the cost of moving people to the city or sending specialists to these areas when we know that they will not work efficiently and will leave. Therefore, there is a need to promote, vulgarise and implement this modality of health care in our context.

Table 3. Comparison of costs for the implementation of tele-ECG by centres

Intervention centres	Type of costs	Costs (F CFA)	Control centres	Type of costs	Costs (F CFA)
Mbouda District Hospital	Variable costs	347 867	**Fombot District Hospital**	Variable costs	0
	Fixed costs	1 434 520		Fixed costs	0
	Total costs	**1 782 387**		**Total costs**	**0**
Akonolinga District Hospital	Variable costs	202 765	**Sa'a District Hospital**	Variable costs	0
	Fixed costs	1 375 520		Fixed costs	0
	Total costs	**1 578 285**		**Total costs**	**0**

References

[1] Damorou F, Baragou S, Pio M, Afassinou YM, N'da N, Pessinaba S, et al. Morbidité et mortalité hospitalière des maladies cardiovasculaires en milieu tropical : exemple d'un centre hospitalier à Lomé (Togo). Pan Afr Med J 2014;17.

[2] Schmidt C, et al. A novel integrated care concept (NICC) versus standard care in the treatment of chronic cardiovascular diseases: protocol for the randomized controlled trial CardioCare MV. Trials. 2018;19:120.

[3] WHO Cameroon (2016). https://www.who.int/nmh/countries/cmr_fr.pdf?ua=1

[4] Kaddar M, Stierle F, Schmidt-Ehry B, Tchicaya A. L'accès des indigents aux soins de santé en Afrique subsaharienne. Tiers Monde 2000;41:903–25.

[5] Bagayoko C.O, et al. Medical and economic benefits of telehealth in low- and middle-income countries: results of a study in four district hospitals in Mali. BMC Health Services Research 2014; 14.

[6] Van Os-Medendorp H, et al. E-health in caring for patients with atopic dermatitis: A randomized controlled cost-effectiveness study of internet-guided monitoring and online self-management training. Br J Dermatol 2012;166:1060–8.

[7] Fatoye F, et al. The Clinical and Cost-Effectiveness of Telerehabilitation for People With Nonspecific Chronic Low Back Pain: Randomized Controlled Trial. JMIR Mhealth Uhealth. 2020 Jun 24;8(6):e15375.

[8] Dixon P, et al. Cost-effectiveness of telehealth for patients with raised cardiovascular disease risk: Evidence from the Healthlines zrandomized controlled trial. BMJ Open 2016;6:1–10.

[9] Bediang G, et al. Evaluation of the Effectiveness of Telemedicine in the Management of Cardiovascular Diseases in Primary Health Care in Cameroon: An Interventional Study. Stud Health Technol Inform. 2021 May 27;281:615-619.

[10] Bogou. http://raft.unige.ch/bogou/, last accessed January. 2022.

[11] Gruska M, Aigner G, Altenberger J et al. Recommendations on the utilization of telemedicine in cardiology. Wiener klinische Wochenschrift. 2020; 132(23): 782-800.

Challenges of Trustable AI and Added-Value on Health
B. Séroussi et al. (Eds.)

doi:10.3233/SHTI220624

Decreased Memory Bias via a Mobile Application: A Symptom Tracker to Monitor Children's Periodic Fever

Helena VALLO HULT[a,b,1], Anna Sigridur ISLIND[c], Karin RYDENMAN[b,d] and Per HÄLLSJÖ WEKELL[b,d]
[a] *University West, School of Business, Economics and IT, Sweden*
[b] *NU Hospital Group, Sweden*
[c] *Reykjavik University, Iceland*
[d] *University of Gothenburg, Sweden*

Abstract. Memory bias, the tendency to rely on certain events over others, can become an issue in chronic illnesses, especially when symptoms are reported retrospectively. This paper examines a case where continuous symptom registration can be facilitated, memory supported, and memory bias reduced by introducing a mobile application. The aim of the paper is to report on the design of an app for collecting subjective data over an extended period to continuously follow children with periodic fever. The research approach is qualitative, building on interview data. The design method is co-design, a collaborative and participatory approach involving researchers, physicians and other key stakeholders, with focus on the views of the parents. We argue that collecting data continuously through an app moves the discussion from memory to the specific data points, which is illustrated through trends shown in the visualizations of the data. Moreover, we highlight the importance of systematically collecting data over an extended period through a data-driven approach to both forward clinical practice and research on complex, often chronic topics such as periodic fever, which is genuinely under-researched to date.

Keywords. Memory bias, healthcare, co-design, children, mobile apps

1. Introduction

Memory bias is a notion that has been widely debated as one of the major biases when it comes to cognitive biases in general [1]. A cognitive bias refers to a systematic error and a simplification in the thinking process that occurs in the brain as people are attempting to interpret the information at hand [2]. More specifically, memory bias refers to the tendency to intentionally or unintentionally rely on recalling certain events and autobiographical memories and favoring those memories over others [3]. This recall process often relates to significant events, including traumatic, unconventional events, or even to systematically selecting the most recent events in a series of events [4]. Memory bias and the tendency to rely on most recent events are especially visible when dealing with a prolonged period such as a chronic condition of any type [5]. Memory

[1]Corresponding Author, Helena VALLO HULT, University West and NU Hospital Group, Trollhattan, Sweden; E-mail: helena.vallo-hult@hv.se.

bias can become an obstacle to giving an adequate description of symptoms, which is crucial in most diagnostic processes as well as follow-up of chronic conditions. It applies particularly to disorders where no specific diagnostic test or marker is available. Periodic Fever, Aphthous Stomatitis, Pharyngitis, and Cervical Adenitis, PFAPA, is an autoinflammatory disorder characterized by regularly recurring fever episodes accompanied by one or more of the features in the acronym. It is generally seen as the most common autoinflammatory disorder among children in many parts of the world. The diagnosis is mainly based on the patient's medical history and symptoms, but there is no specific test to confirm or exclude the condition. Instead, it is necessary to follow the course of the disease over time [6,7]. A symptom tracker is a valuable tool for tracking the disease progression, characterizing the fever episodes, and getting an overview of their regularity. However, it can be challenging for the parents to document and report symptoms in a structured manner. This paper examines a case where memory bias can be reduced drastically by introducing a mobile application (an app) for the system tracker. The app has three primary functions: i) to be readily available for the parents to register symptoms regularly; ii) to visualize data over time to reduce memory bias and; iii) to feed data to the pediatricians treating the children. This type of continuous data collection could improve how children with periodic fever can be cared for, clinically assessed, and how the condition can be researched. The paper report on the co-design of an app to support memory, facilitate continuous data collection, and reduce memory bias for children with periodic fever. We focus on the parents as valuable stakeholders in the co-design process.

2. Related Research

PFAPA syndrome typically has an onset under the age of 5 and is a common and important differential diagnosis among preschool children with recurrent fevers [6, 7]. The fever episodes usually last 3-6 days and typically recur regularly with an interval of 3-6 weeks [6]. Between the fever episodes, the patient usually is symptom-free with normalized inflammatory variables and grows normally. Awareness and recognition of PFAPA are vital for providing adequate treatment and avoiding misdiagnosis [6, 7]. While mHealth apps have benefits such as accessibility, cost reduction, improved patient quality of life, alongside more precise and personalized disease management, to fulfill the positive outcomes, they need to be linked to clinical practice and adapted to both patients and healthcare [8]. A symptom tracker can be used as a way to gather patient-generated health data (PGHD), which can help patients organize details and activities of the illness, increase engagement, and promote patient empowerment while enhancing adherence to treatment [9, 10]. Factors, such as differences in parental approach to fever management and methods of diagnosis may also influence the design of the tracker and whether memory bias is decreased [11]. It is important to involve relevant stakeholders in the design process to reach a fruitful design. The fundamentals of co-design entail that the relevant stakeholders have a voice in the design processes that will ultimately affect their lives [12, 13]. Including tertiary stakeholders, such as parents, in co-design can lead to digital technology that will both be relevant and useful [14, 15]. In co-design, all stakeholders that partake in the co-design process are regarded as intelligent, active partners who actively contribute to the design process through their insights into their expertise (in this case their expertise from taking care of their children with PFAPA), an expertise that will be shaped by the technology that is being designed significantly.

3. Methods

In this paper, we report on an ongoing project and the combined experiences of 100 parents that have contributed to our co-design process. In this phase, we wanted a broader perspective and therefore did not limit participants only to parents to children with periodic fever. The app will be used in Sweden to support parents of children with PFAPA systematically and continuously digitally log and track symptoms. The multidisciplinary research team consists of researchers and designers with pediatrics and health informatics expertise. The research approach is qualitative, and the data collection for this phase was conducted through semi-structured interviews with parents. All interviews were read several times, collaboratively coded, and analyzed through content analysis [16], and validated with pediatricians. As mentioned above, the design approach is co-design, and the project considers all stakeholders involved in the co-design process. In contrast, this paper focuses on the views of the parents (based on the interview data) and implications when co-designing for continuous monitoring of children.

4. Results

Although still in the design phase, we already see that an app for tracking children's fever is essential and that parents are critical stakeholders due to their vast experience of caring for their children. For instance, regarding having margins set in the app with a specific fever level that could indicate when to contact a healthcare professional, one parent said: *"It might be good to have, just to be on the safe side if there are any uncertainties and so on, but I am rarely uncertain."* Another parent corroborated: *"It might be ok, but I do know it so well myself, from years of experience."* Yet another parent thought that others might need it, but not them: *"It might not be so bad to include for inexperienced people but me, I don't need it, I know it."* These quotes clarify that the parents are the experts, know their children, and have instincts that they rely on wholeheartedly. We conclude that parents are essential stakeholders in a co-design process that aims to aid their care for their children. Next, we summarize the findings, focusing on design features in the app that facilitate reflections, increase compliance and reduce memory bias.

I. Notifications to increase compliance. Regarding notifications, one parent uttered: *"O, yes. That and just that would help me enormously."* Also, the same parent said: *"It would be an easy way to track the information and to have an overview of the information too. Everything is always there"* Another parent shared a similar opinion: *"App, yes, definitely…I would lose the paper and forget it at home when I would go to the doctor with my child."* A third parent offered the opinion that notifications might even help with the systematics in the registration: *"It is good to have something that helps you remember the important things, and it would help to systematically register the symptoms every day at the same time."* Yet another parent expressed that it could help during stressful times: *"I might forget, when my child is sick I am sometimes going on very little sleep, but if I would get a notification, then I would still register. So, notifications would help me."* The parents would gain from having notifications implemented within the app.

II. Visualizations as a crucial factor for reflection. Regarding how the information registered in the app could be presented, one parent suggested: *"Maybe some kind of graph that shows a clear and simple pattern in the information. Perhaps we should try*

to fill in more information than symptoms to see if there is any correlation between external factors and changes in the body. That would be really good," while another parent said: *"Just simple. The only thing that matters is that I am able to see a trend for a specific time period. Everything needs to work fast and efficiently. I rarely have time that I can afford to waste when I have a sick child."* One parent also offered the view that the visualizations might help with reflection: *"If I could see how it is over time, then I could also see when it is better because sometimes it might be better for a longer time as well, but it is difficult to know without seeing it visually."* From these quotes, it is apparent that visualizations are a crucial part of an app for monitoring periodic fever.

III. Asynchronous messages to decrease the risk of forgetting to notify changes. Several parents reported that they wanted to communicate directly and securely with their pediatrician within the app. The messages, or chat, would have to be asynchronous, as the parents are aware that they might want to send a message to their pediatrician in the middle of the night. Still, they do not expect their pediatrician to answer them. They do mind that the messages are only delivered in the morning since they want the messages within the app only because they are afraid to forget to send the message if they are unable to do it when they come to think of it, due to extended sleep deprivation. On that note, one parent described: *"Sometimes, when my child is sick, I wake up, and it might be days since I got a good night's sleep. Those nights, it would be good to be able to send a message just there and then, not because I want a reply there and then, but because I really don't want to forget to send the pediatrician a message if something has drastically changed."* Sending messages in an asynchronous format would be helpful to reduce the risk of the parents forgetting to send them.

5. Discussion

In this section, we discuss the design of the app, focusing on the parents' views and how specific app features can help with compliance while reducing memory bias [3-5].

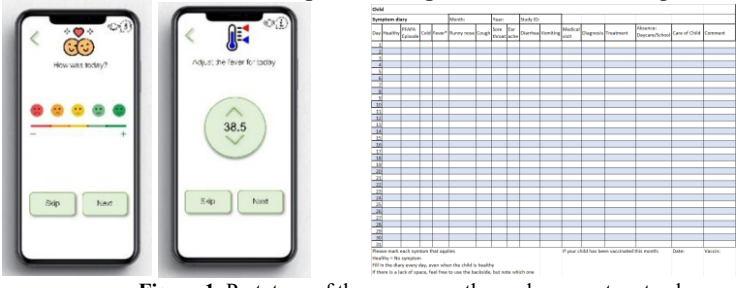


Figure 1. Prototype of the app versus the analog symptom tracker used before

The app (see Figure 1) has three primary goals: i) to be readily available for the parents to register symptoms regularly; ii) to visualize data over time to reduce memory bias and; iii) to feed data to the pediatricians treating the children. The main findings drawn from the co-design process [14,15] comprise features that our app needs, and features which we hereby recommend for others that are designing mHealth apps to aid parents in caring for their children: i) ***notifications*** to increase compliance, ii) ***visualizations*** as a crucial factor for reflection, iii) ***asynchronous messages*** to decrease the risk of forgetting to notify about changes. Combined, these features should certainly contribute to decreased

memory bias and increased compliance. Even if this first part of the design process did not specifically include parents to children with PFAPA, we believe that a user-friendly and co-designed app can be a valuable tool in diagnosing this disorder as well as in assessing the symptoms over time. If used in a wider context, we also believe that the symptom tracker can create awareness and recognition of PFAPA, as the regularity of fever episodes could be more reliably documented [6,7]. We plan to continue by involving parents and pediatricians in the co-design process to develop a useful symptom tracker and connecting the app to the context of the patients and healthcare practice [8].

6. Conclusion

This paper suggests that the use of apps may reduce memory bias. Collecting data continuously, through an app, moves the discussion from memory to the specific data points, illustrated through trends shown in the data visualizations. Moreover, the study underlines that parents are essential stakeholders and should be included in the design process. Our findings highlight the importance of systematically collecting data over an extended period through a data-driven approach to forward clinical practice and research on complex, chronic topics such as periodic fever, which is under-researched to date.

References

[1] Harding EJ, Paul ES, Mendl M. Cognitive bias and affective state. Nature. 2004;427(6972):312-.
[2] Enke B, Gneezy U, Hall B, Martin DC, Nelidov V, Offerman T, et al. Cognitive Biases: Mistakes or Missing Stakes? : National Bureau of Economic Research; 2021.
[3] Bianchi R, Laurent E, Schonfeld IS, Bietti LM, Mayor E. Memory bias toward emotional information in burnout and depression. Journal of Health Psychology. 2020;25(10-11):1567-75.
[4] Mogg K, et al. Memory bias in clinical anxiety. Journal of abnormal psychology. 1987;96(2):94.
[5] Lindroth T, Islind AS, Steineck G, Lundin J, editors. From Narratives to Numbers: Data Work and Patient-Generated Health Data in Consultations. MIE; 2018.
[6] Wekell P, et al. Review of autoinflammatory diseases, with a special focus on periodic fever, aphthous stomatitis, pharyngitis and cervical adenitis syndrome. Acta Paediatrica. 2016;105(10):1140-51.
[7] Rydenman K, Berg S, Karlsson-Bengtsson A, Fasth A, Wekell P. PFAPA-syndrom–en viktig differential-diagnos hos barn med återkommande feberepisoder [PFAPA syndrome – An important differential diagnosis in children with recurrent fever]. Lakartidningen. 2019;116:FP9U.
[8] Peyroteo M, Ferreira IA, Elvas LB, Ferreira JC, Lapão LV. Remote Monitoring Systems for Patients With Chronic Diseases in Primary Health Care: Systematic Review. JMIR Mhealth Uhealth. 2021;9(12):e28285.
[9] Islind AS, Vallo Hult H, Johansson V, Angenete E, Gellerstedt M. Invisible Work meets Visible Work: Infrastructuring from the Perspective of Patients and Healthcare Professionals. HICSS, January 5-82021.
[10] Islind AS, et al. Individualized blended care for patients with colorectal cancer: the patient's view on informational support. Support Care Cancer. 2020.
[11] Villarejo-Rodríguez MG, Rodríguez-Martín B. Parental Approach to the Management of Childhood Fever: Differences between Health Professional and Non-Health Professional Parents. International Journal of Environmental Research and Public Health. 2019;16(20).
[12] Joshi SG, Bratteteig T. Designing for prolonged mastery. On involving old people in participatory design. Scandinavian Journal of Information Systems. 2016;28(1).
[13] Kensing F, Greenbaum J. Heritage: Having a say. Routledge international handbook of participatory design: Routledge; 2012. p. 41-56.
[14] Vallo Hult H, Islind AS, Master Östlund C, Holmgren D, Wekell P. Sociotechnical Co-design with General Pediatricians: Ripple Effects through Collaboration in Action. AMCIS; Salt Lake City, UT, USA 2020.
[15] Islind AS, Lundh Snis U. From co-design to co-care: designing a collaborative practice in care. Systems, Signs & Actions. 2018;11(1):1-24.
[16] Graneheim UH, Lundman B. Qualitative content analysis in nursing research: concepts, procedures and measures to achieve trustworthiness. Nurse education today. 2004;24(2):105-12.

Challenges of Trustable AI and Added-Value on Health
B. Séroussi et al. (Eds.)

doi:10.3233/SHTI220625

Dementia-Related Barriers to mHealth Use: Validation by Expert Opinion

Thomas ENGELSMA[a,b,1], Ahsen YURT[b], Monique W. JASPERS[b] and Linda W. PEUTE[b]
[a] *Center for Human Factors Engineering of Health Information Technology*
[b] *Department of Medical Informatics, Amsterdam Public Health research institute, Amsterdam UMC, location AMC, Amsterdam, The Netherlands*

Abstract. mHealth use for people with dementia is fraught with factors influencing its implementation in care and daily life. A better understanding of these factors may provide guidelines to inclusive design. This study aimed to assess whether factors gathered in a literature-based model could be validated by opinions of experts. On basis of a questionnaire as part of a larger study, experts identified barriers that they considered to be related to aging and dementia influencing mHealth use. Nineteen barriers that were mentioned by the dementia experts were covered in our literature-based model. No adaptions to the model were required. The dementia experts acclaimed three barriers to mHealth use that could not be mapped onto the framework: the unavailability of (informal) caregivers to support the mHealth use, the stage and type of dementia of an mHealth user, and the fear of the unknown. These should be considered as prerequisites in the implementation phase of mHealth and explored more in future research.

Keywords. Dementia, Mobile Health, Design, Validation

1. Introduction

Digital health technologies (DHT) use for a wide range of end-users is increasing. To enable successful implementation, adoption, and retention of DHT for a wide range of end-users, it is important to ensure inclusive design and keep the needs, capabilities, and barriers of these end-users with respect to DHT use into account. Previously, a literature-based model has been developed capturing barriers to DHT use, more specifically to mHealth use, for older adults with dementia [1]. This model is based on a framework, MOLD-US, that defines general aging barriers to mHealth use [2] and extended with dementia-related barriers identified in clinical literature and usability studies of mHealth performed with older adults with dementia. To validate this model, from now on referred to as MOLDEM-US, a larger study is ongoingwith dementia experts to prioritize the actual impact of these barriers to mHealth use by these adults and how often these occur in practice. As part of this study, we gathered qualitative data with regards to these experts' experiences with barriers to mHealth use for these adults. With these data, we

[1] Corresponding Author, Thomas Engelsma, Amsterdam UMC, University of Amsterdam, Department of Medical Informatics, Center for Human Factors Engineering of Health Information Technology, Meibergdreef 9, Amsterdam, the Netherlands; E-mail: t.engelsma@amsterdamumc.nl

aim to perform an additional validation and potential refinement of MOLDEM-US based on real-life practice.

2. Methods

We performed a larger study with dementia experts (data gathered between April 12th 2021 till June 22th 2021) to prioritize the actual impact and occurrence of barriers to current mHealth use for older adults with dementia. The recruited dementia experts consisted of case managers, district nurses, hospital healthcare workers, informal caregivers, and researchers. Recruitment took place via various organizations related to dementia care and through snowball sampling. An introductory questionnaire was submitted to gain insights into the experience of dementia experts with barriers to DHT use for older adults with dementia. For this paper the responses of dementia experts to the question: "Which barriers related to aging and dementia influence the use of mHealth?" were analyzed. The questionnaire was distributed using Castor EDC [3]. The barriers appointed in the qualitative responses were extracted from the data and mapped, using a thematic analysis, on the barriers captured in MOLDEM-US related to cognition, perception, physical ability, frame of mind, or speech- and language. While completing the introductory questionnaire, participants were not aware of the barriers in MOLDEM-US and were asked to answer the questions based on their own experiences. Initial analysis was performed by AY, and this was adapted by TE and LWP until consensus was reached.

3. Results

3.1. Participant characteristics

Fifty-four dementia experts were invited to participate in the questionnaire of which 48 responded. However, an additional four participants were excluded due to no meaningful response to the open-ended questions of the questionnaire. With a response rate of 81%, 44 participants were included in the data analysis to identify barriers to mHealth use. These participants were mainly female (90.9%), smartphone users (100%) and reported as main function: case manager (n=17), hospital or district nurse (n=15), geriatrician (n=1), researcher (n=5), informal caregivers (n=4) or project coordinator (n=2). Finally, 65% mentioned to have experience with mHealth use.

3.2. Validation of the model

Nineteen barriers already included in MOLDEM-US were identified through qualitative data from our survey (Table 1). Most barriers were validated in the "frame of mind barrier" category (n=7), followed by "cognitive barriers" (n=6), "perception barriers" (n=3), "physical ability barriers" (n=2), and "speech- and language barriers" (n=1). Additionally, three barriers identified through the questionnaire did not match with any of the barriers in MOLDEM-US and could therefore not be mapped: 'no availability of

(informal) caregiver to support mHealth use', 'the fear for the unknown', and 'the stage and type of dementia'.

Table 1. Validation of barriers from the extended MOLD-US model by dementia experts order by most mentioned.

Barriers from extended MOLD-US model	n	Quotes from questionnaire
Cognitive barriers		
Learnability	13	*"Can people with dementia still learn new things such as the use of mHealth?"*
Working memory	11	*"The decline in cognition affects the processing of the large amount of information that mHealth asks of someone"*
Recognition	6	*"Ability to remember actions for a longtime."*
Ability to organize thoughts or actions	3	*"Forgetfulness, forgetting passwords, not knowing where something is stored."*
Reasoning	2	*"Dementia makes it more difficult to distinguish between cause and effect"*
Making decisions and judgements	1	*"Little problem-solving ability. For example, as soon as the WiFi freezes, panic and mistrust arises in the functions."*
Physical ability barriers		
Hand-eye coordination	3	*"In Parkinson's (dementia) the (fine) motor skills fail. So you can no longer operate it. Operating the telephone was one of the first things my husband could no longer do."*
Tremor	3	*"But also, for example, tremors in parkinsonism can cause difficult situations when using a telephone or tablet."*
Perception barriers		
Color vision	2	*"Dark colors"*
Touch sensation	1	*"Difficulty with buttons/touch screens"*
Ability to say what has been seen	1	*"Delusions and/or visual hallucinations may be enhanced"*
Frame of mind barriers		
Computer literacy	13	*"If people with dementia themselves have to do something with a tablet or mobile phone, it often becomes difficult, because they did not grow up with it. Keep in mind that this might be easier for future generations."*
Trust in own ability	5	*"People [with dementia] dread learning it and think they can't"*
Efficiency in seeing benefits	5	*"Most older people come from a generation with little technology. Older people with dementia are therefore even less able to understand how mHealth can be used and what its benefits are."*
Perceived complexity	5	*"Dealing with technology can quickly become overwhelming."*

Fluctuating / Concentration issues attention	3	*"People are easily distracted from possible pop-ups, or advertisements in apps"*
Restlessness and agitation	2	*"Can't be patient anymore and get frustrated faster."*
Stigmatization	1	*"They themselves do not notice what is going wrong/different than before."*
Speech- and language barriers		
Semantic knowledge	2	*"Difficulty using and functioning of a smartphone, etc. due to difficulty with the language used"*

4. Discussion

This paper sought to gain insights into known barriers to mHealth use for older adults with dementia according to dementia experts. Nineteen barriers were identified and already covered in MOLDEM-US, which supports its validation and no additional refinements to the model were necessary. According to the dementia experts in this study, a barrier to mHealth use for older adults with dementia is the unavailability of (informal) caregivers to support the mHealth use. It can be implied that this barrier is rather a prerequisite for older adults with dementia to use mHealth. The need for support can be traced back to other frequently mentioned barriers by the dementia experts such as difficulties with learning new skills (*learnability*), *perceived complexity*, *trust in own abilities*, and low *computer literacy*. One or more of these four barriers were mentioned by 59% of participants.

A second acclaimed barrier which can be considered a prerequisite to mHealth use is the stage and type of dementia of an mHealth user. The current version of MOLDEM-US model does not discriminate on the severity of barriers per stage of dementia, but found that participants in usability studies were mostly diagnosed with mild cognitive impairment or early-mild dementia [1]. However, it is relevant for future research to investigate how inclusive design approaches can contribute to increase the durability of mHealth technologies as the disease progresses. This is in line with findings from a scoping review by Guzman-Perra et al. who state durability is a factor that can influence the adoption of smart health technology for people with dementia [4].

A third barrier and final prerequisite to mHealth use, "the fear of the unknown" was mentioned by few participants. The interpretation of this barrier is rather broad, but might relate to barriers in MOLDEM-US such as *trust in own abilities* and *computer literacy.* Previous research suggests that older adults who do not use digital devices may experience computer-related anxiety which makes them feel technophobic or unconfident in using digital solutions [5].

Finally, the MOLDEM-US barriers *learnability, working memory,* and *computer literacy* were most often mentioned by experts. It is important to acknowledge that this does not indicate the frequency of such barriers in practice, but rather the best-known barriers by the dementia experts in this study. However, in the larger Delphi study that is being performed by the authors we aim to prioritize these barriers in contribution to the development of design guidelines supporting future mHealth innovations.

A strength of this study is the inclusion of a broad range of dementia experts with various experiences with older adults with dementia, but also various settings in which

mHealth is or can be implemented. Another strength is the initial and secondary analysis that have been performed by three researchers to reach consensus on the mapping of the responses to the known barriers. A limitation might be that not all experts reported to have mHealth experience. However, their experience with working with older adults with dementia provided valuable responses to the questionnaire as mentioned barriers were covered in our model.

Future work consists of further investigating the impact of barriers to mHealth use and the development of inclusive design suggestions for the barriers that most affect current mHealth use in order to improve the implementation, adoption, and retention of mHealth in dementia care.

5. Conclusions

Barriers covered in MOLDEM-US influence mHealth use according to the experiences of dementia experts in this study. This ensured no additional refinements to the model were required. Three barriers were acclaimed to influence mHealth use, but should be considered as prerequisites for the actual implementation of mHealth. These should be accounted for during implementation, to tackle barriers such as learnability, perceived complexity, trust in own abilities, and computer literacy. Future work needs to provide practical insights for mHealth developers to contribute to inclusive design for older adults with dementia.

References

[1] Engelsma T, Jaspers MWM, Peute LW. Considerate mHealth design for older adults with Alzheimer's disease and related dementias (ADRD): A scoping review on usability barriers and design suggestions. Int J Med Inform. 2021 Aug;152:104494.

[2] Wildenbos GA, Peute L, Jaspers M. Aging barriers influencing mobile health usability for older adults: A literature based framework (MOLD-US). Int J Med Inform. 2018;114:66-75.

[3] Castor EDC. (2019). Castor Electronic Data Capture. [online] Available at: https://castoredc.com.

[4] Guisado-Fernández E, Giunti G, Mackey LM, Blake C, Caulfield BM. Factors Influencing the Adoption of Smart Health Technologies for People With Dementia and Their Informal Caregivers: Scoping Review and Design Framework. JMIR Aging 2019;2(1):e12192

[5] Di Giacomo D, Ranieri J, D'Amico M, Guerra F, Passafiume D. Psychological Barriers to Digital Living in Older Adults: Computer Anxiety as Predictive Mechanism for Technophobia. Behav Sci (Basel). 2019 Sep 11;9(9):96. doi: 10.3390/bs9090096. PMID: 31514364; PMCID: PMC6770433.v

Challenges of Trustable AI and Added-Value on Health
B. Séroussi et al. (Eds.)

doi:10.3233/SHTI220626

Differentiation of Cell Painted Organelles Using Non Local Texture Descriptor and Random Forest Approach

Sreelekshmi PALLIYIL SREEKUMAR [a,1], Rohini PALANISAMY[b] and Ramakrishnan SWAMINATHAN[a]
[a] *Biomedical Engineering Group, Centre of Excellence in Medical Device Regulations and Standards, Department of Applied Mechanics, Indian Institute of Technology Madras, Chennai, Tamil Nadu, India*
[b] *Indian Institute of Information Technology, Design and Manufacturing, Kancheepuram, Chennai, Tamil Nadu, India*

Abstract. Discriminating the cell organelles from microscopic images is a challenging task due to their high similarity in image appearance. In this work, an attempt has been made to differentiate nuclei, Endoplasmic Reticulum (ER) and cytoplasm using a texture pattern descriptor and Random Forest classifier. For this, Cell Painted public dataset from Broad Bioimage Benchmark collection are considered. Texture features are extracted from each image using Non Local Binary Pattern (NLBP) that captures the relationship between global pixels and sampling instances in a local neighborhood. Non local central pixels called anchors are derived from central pixels of image patches and compared with sampling instances. Binary string generated from this is encoded into 29 patterns. Statistical one-way analysis of variance (ANOVA) is performed to select significant features and are validated using Random Forest classifier. The dependency of classifier performance on the local patch radius (R) and the number of anchors (K) are also evaluated. The results indicate that 8 patterns out of 29 are showing strong inter class variability with high F value. Classification accuracy of 84% is achieved with R=3 and K=5. Experimental results demonstrate that the proposed work captures complex patterns in cell structure useful for differentiating cell components which can be employed for evaluating the cytotoxic effects in cell lines.

Keywords. Cell Painting, Non Local Binary Pattern, Texture Feature, Random Forest

1. Introduction

Cell Painting is an advanced imaging technique that uses different fluorescent probes to target specific cell organelles such as nuclei, Endoplasmic Reticulum (ER), cytoplasm, mitochondria and Golgi apparatus for profiling subtle patterns in cell structure [1]. From the previous studies it can be seen that textures of different cell components are different

[1] Corresponding Author, Sreelekshmi Palliyil Sreekumar, Biomedical Engineering Group, Centre of Excellence in Medical Device Regulations and Standards, Department of Applied Mechanics, Indian Institute of Technology Madras, Chennai, Tamil Nadu, India; E-mail: lekshmishyam2009@gmail.com.

[2]. Hence the spatial relationship in the Cell Painted microscopic images can be effectively measured by texture patterns.

Local binary descriptors have received wide acceptance in characterizing the local features in an image [3]. Local binary pattern (LBP) and its derivatives focus on the spatial relation between central pixel and sampling instances in the local neighborhood [4]. These methods are inadequate in describing the long range pixel interaction that occurs outside a compact region which can also be considered as important for feature representation. In this work, wide range pixel relationship is captured by Non Local Binary Pattern (NLBP) based on global image statistics rather than local connected region in the image [5].

2. Methods

The proposed methodology for categorizing the cell organelles is described in Figure1.

2.1 Image Dataset

For this study, Cell Painted images of Human U2OS cells are obtained from Broad Bioimage Benchmark Collection [6]. 3456 images of nuclei,3456 images of ER and 3456 images of cytoplasm in 16-bit grayscale of size 1080x1080 pixels are considered.

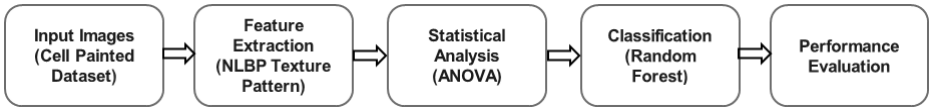


Figure 1. Block diagram of proposed Methodology

2.2 Non Local Binary Pattern

To extract texture features, a local image patch of size ω x ω with central pixel x is considered and their gray values are sorted in ascending order as given in the Eq. (1)

$$\bar{g}'_{c1}, \ldots\ldots, \bar{g}'_{cN} := \text{sort}(\bar{g}'_{c1}, \ldots\ldots, \bar{g}'_{cN}) \quad (1)$$

where $\bar{g}'_{cN}$ represents the gray value of the N^{th} sorted central pixel and N is the total number of central pixels [5]. The sorted pixels are then divided into K equal intervals and the anchors are calculated for each interval using Eq. (2)

$$g_{Ak} = \frac{1}{\lfloor N/K \rfloor} \sum_{n=(k-1)\lfloor N/K \rfloor+1}^{k\lfloor N/K \rfloor} \bar{g}'_{cN} \quad (2)$$

Where g_{Ak} (1,…,K) represents the gray value of the k^{th} anchor and $\lfloor . \rfloor$ is the floor function. The center pixel and its neighbors are shown in Figure 2 by a yellow rectangle and anchors are computed for K=2. Each of these anchor values are compared with sampling instances of center pixels and binary pattern is generated based on given $s(x)$ condition.These binary strings are further encoded into distinct NLBP codes for different R and P values based on Eq. (3).In this work, R=3 and P=24 are considered and hence 29 NLBP codes are obtained according to U value given by Eq.(4)

$$NLBP_{R,P,K} = \begin{cases} \sum_{p=0}^{P-1} s(\bar{g}_{R,P} - g_{Ak}) & , U(NLBP) \le 2 \\ P+1 & , U(NLBP) = 4 \\ P+2 & , U(NLBP) = 6 \\ P+3 & , U(NLBP) = 8 \\ P+4 & , U(NLBP) = 10 \\ P+5 & , Otherwise \end{cases} \quad (3)$$

where R is the local patch radius, P is the number of sampling instances, K is the number of anchors and U is a uniformity measure. U is expressed as

$$U(NLBP) = \left|s(\bar{g}_{R,P-1} - g_{Ak}) - s(\bar{g}_{R,0} - g_{Ak})\right| + \sum_{p=1}^{P-1} \left|s(\bar{g}_{R,P} - g_{Ak}) - s(\bar{g}_{R,p-1} - g_{Ak})\right|$$

$$\text{where } s(x) = \begin{cases} 1, x \ge 0 \\ 0, x < 0 \end{cases}, \; x = \sum_{p=0}^{p-1} (\bar{g}_{R,P} - g_{Ak}) \quad (4)$$

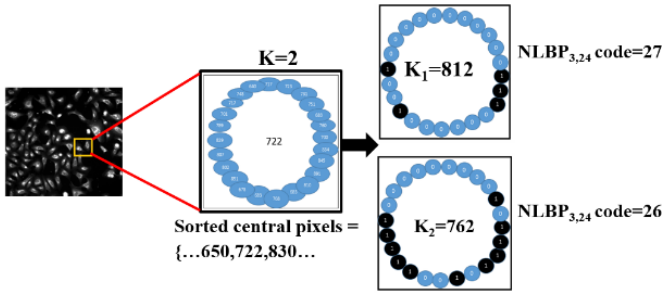


Figure 2. Demonstration of NLBP code generation

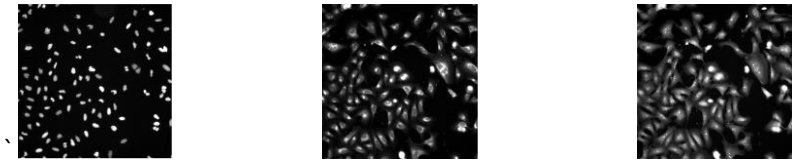
Statistical one-way analysis of variance (ANOVA) is carried out on these NLBP codes to select significant patterns and are fed to the Random forest classifier.

2.3 Classification

Random Forest algorithm is a widely used artificial intelligence technique in medical data classification due to their feature ranking and selection methodology on a random split basis. It splits the feature vectors into different sample sets and builds multiple decision trees with randomly selected features. Each decision tree produces result and finally, the average of multiple tree predictions is taken for decision making [7].

3. Results and Discussion

The representative images of nuclei, endoplasmic reticulum and cytoplasm are shown in Figure 3. It is observed that subtle differences between these microscopic images cannot



(a) (b) (c)

Figure 3. Representative input images (a) Nuclei, (b) Endoplasmic Reticulum, (c) Cytoplasm

be accessed on visual examination. The box plots shown in Figure 4 (a) and (b) are the energy distribution of representative patterns. It is evident from the box plots of patterns 23,25,26 and 27 that there exists a wide difference between the interquartile range and median values of given organelles. Therefore, the spatial heterogeneity between nuclei, ER and cytoplasm are uncovered with these texture measures.

The extracted NLBP features are analyzed for statistical significance using ANOVA. The patterns 1, 22,23 and 25 to 29 show higher F value which indicates strong inter class variability among cell organelles. It is observed that except pattern 1, majority of significant ($p < 0.05$) patterns capable of discriminating nuclei, ER and cytoplasm have uniformity measure greater than 2. This shows that NLBP quantifies non uniform patterns which corresponds to complex textures such as high curvature edges, lines and corners.

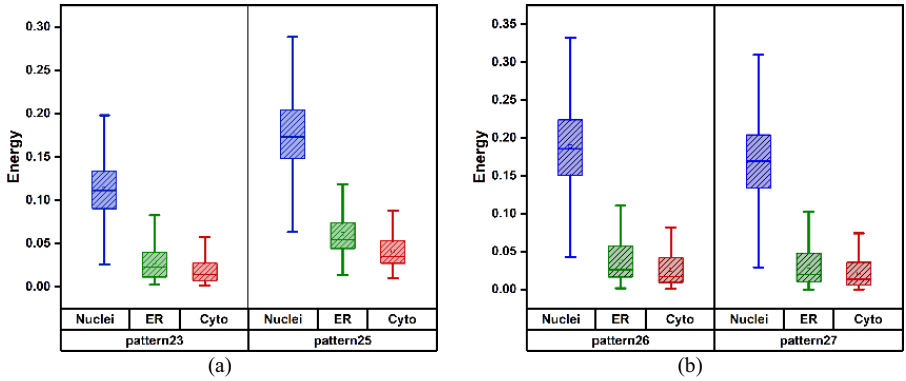
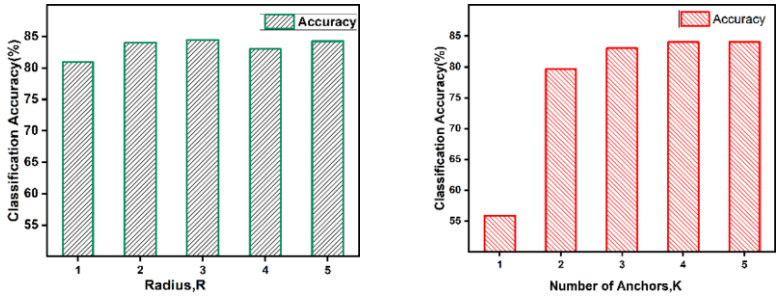


Figure 4. (a) Box plots of patterns 23 & 25 ,(b) Box plots of patterns 26 & 27

The effect of R and K on classification accuracy are shown in Figure 5(a) and (b). This method achieved an accuracy of 84% with R=3 and K=5.With increase in number of anchors K, NLBP progressively captures intensity variations of local patch with respect to the whole image. The increase in accuracy with increase in R and K is associated with the ability of NLBP to comprehensively capture the underlying spatial relationship in the microscopic images.

Pawlowski N et al. [8] had proposed pre trained neural networks such as ResNet-152, VGG 16, Inception-v3 for extracting meaningful features from cell painted images for classification and obtained an accuracy of 55.34%,66.02% and 70.87% respectively. Goldsborough P et al. [9] had presented Least Squares Generative Adversarial Network (LSGAN) for classification of mechanism of action of chemicals and achieved 68% accuracy. It is observed that the proposed method shows better performance with an accuracy of 84% than the state-of-the art methods. This indicates the clinical relevance of this study in Artificial Intelligence (AI) based solutions for problems in health informatics.



(a) (b)

Figure 5. (a) Effect of local patch radius on the classification accuracy, (b)Effect of number of Anchors on the classification accuracy

4. Conclusions

In this study, differentiation of cell painted organelles using a Non Local Texture Descriptor with Random Forest classifier is performed. The results demonstrate that NLBP captures non uniform patterns that occur due to wide range pixel interaction in large neighborhoods. This method illustrates the feasibility of global image statistics for effective categorization of cell components in microscopic images. Proper selection of local patch radius and number of anchors are important for accurate classification. This work could be useful for analyzing the effect of cytotoxicity and understanding the reaction of chemical compounds in cell lines. The proposed machine learning based study aids in analyzing the cytological effects with minimum medical expertise.

References

[1] Bray MA, Singh S, Han H, Davis CT, Borgeson B, Hartland C, Kost-Alimova M, Gustafsdottir SM, Gibson CC, Carpenter AE. Cell Painting, a high-content image-based assay for morphological profiling using multiplexed fluorescent dyes. Nature protocols. 2016 Sep;11(9):1757-74.

[2] Fekri-Ershad S. Cell phenotype classification using multi threshold uniform local ternary patterns in fluorescence microscope images. Multimedia Tools and Applications. 2021 Mar;80(8):12103-16.

[3] Yin H, Chen Y, Xiong J, Xia R, Xie J, Yang K. An improved local binary pattern method for pollen image classification and recognition. Computers & Electrical Engineering. 2021 Mar 1;90:106983.

[4] Liu L, Chen J, Fieguth P, Zhao G, Chellappa R, Pietikäinen M. From BoW to CNN: Two decades of texture representation for texture classification. International Journal of Computer Vision. 2019 Jan;127(1):74-109.

[5] Song T, Feng J, Luo L, Gao C, Li H. Robust texture description using local grouped order pattern and non-local binary pattern. IEEE Transactions on Circuits and Systems for Video Technology. 2020 Feb 6;31(1):189-202

[6] Singh S, Wu X, Ljosa V, Bray MA, Piccioni F, Root DE, Doench JG, Boehm JS, Carpenter AE. Morphological profiles of RNAi-induced gene knockdown are highly reproducible but dominated by seed effects. PloS one. 2015 Jul 21;10(7): e0131370

[7] Alam MZ, Rahman MS, Rahman MS. A Random Forest based predictor for medical data classification using feature ranking. Informatics in Medicine Unlocked. 2019 Jan 1;15:100180.

[8] Pawlowski N, Caicedo JC, Singh S, Carpenter AE, Storkey A. Automating morphological profiling with generic deep convolutional networks. BioRxiv. 2016 Jan 1:085118.

[9] Goldsborough P, Pawlowski N, Caicedo JC, Singh S, Carpenter AE. CytoGAN: generative modeling of cell images. BioRxiv. 2017 Jan 1:227645.

Challenges of Trustable AI and Added-Value on Health
B. Séroussi et al. (Eds.)

doi:10.3233/SHTI220627

Timeliness of Single-Patch 12-Lead Electrocardiography for Patients with Chest Pain at the Emergency Department

Suyoung YOO[a,1], Hansol CHANG[a,b], Taerim KIM[b], Hee YOON[b], Sungyeon HWANG[b], Taegun SIN[b], Minsub SIM[b], Ikjoon JO[b], Jinho CHOI[b] and Won Chul CHA[a,b,c]

[a] *Department of Digital Health, Samsung Advanced Institute for Health Science & Technology, Sungkyunkwan University, Seoul, Republic of Korea*

[b] *Department of Emergency Medicine, Samsung Medical Center, Seoul, Republic of Korea*

[c] *Health Information and Strategy Center, Samsung Medical Center, Seoul, Republic of Korea*

Abstract. A 12-lead ECG is used in emergency departments to diagnose and treat patients with chest discomfort. Performing ECGs at the proper time has been found to increase treatment outcomes. A timer on a wearable ECG ensures proper recording. We compared the timing accuracy of single-patch 12-lead ECG to conventional ECG, expecting the former to be more accurate. Adult patients with chest pain but not in shock were randomized into two groups: SP-ECG and C-ECG. The final analysis included 33 of the 36 recruited patients. The key result was the time taken to record the ECG in both groups. The two groups' average ages were 63.7 and 58.1 years. The SP-ECG group was 87.5 percent timely, while the C-ECG group was 47.0 percent (P.74). At the second follow-up, it was 75.0 percent and 35.2%. Continuous ECG without interfering with other exams is feasible in complex ER circumstances. But the accuracy of single-patch ECG has not been verified. The device also had some minor difficulties. The use of SP-ECG may help alleviate overcrowding or staffing issues in EDs, although more research is required.

Keywords. Electrocardiogram, Wireless Technology, emergency department

1. Introduction

Twelve-lead electrocardiography (ECG) is an essential process in the emergency department (ED) for patients with complaints of chest pain. The most important step in a patient who complains of chest pain is to identify the location of the pain. ECG should be performed to determine if the pain is caused by cardiovascular disease. Timely ECG is associated with improved clinical outcomes in patients with cardiovascular disease. Therefore, delayed ECG in the ED, which can be aggravated when the ED is overcrowded or short-staffed, can result in poor outcomes. [1-2] Complex and unstable circumstances in the ED are challenging for current monitoring systems. Patients are often moved from one place to another for various tests and procedures. In addition, long

[1] Corresponding Author, Su Young Yoo; E-mail: rueseo92@gmail.com.

physical lines, large sizes, and multiple patches of current ECG devices are not suitable for long-term use and require a long preparation time. A single-patch wireless 12-lead ECG (SP-ECG) with a timer could be beneficial in cases of pre-decided follow-up ECG. A single patch would make it possible for patients to move outside the bed and could be used in complex emergency room situations requiring many tests. However, to our knowledge, there have been no studies on such devices in the ED setting. This study aimed to evaluate the effect of a SP-ECG with a timer on the timeliness of follow-up ECG for chest pain patients in the clinical setting.

2. Methods

The study was conducted in an academic tertiary hospital's ED. Participants were divided into two groups: C-ECG and SP-ECG. The key comparative variable was the timing of the two ECG types. The hospital's institutional review board accepted the study protocol (#2019-01-046-008).

The study took place in an academic tertiary hospital in Seoul with 2,000 inpatient beds and 2 million outpatient visits each year. Annually, 78,000 people visit the ED. The study began on July 30, 2020, and ended on October 8, 2020. The trial lasted about 70 days. The study comprised patients who came to the hospital's ED with chest pain. The inclusion criteria were: admissions to the ER with chest pain or discomfort, age over 19, and consent to participate. Shock or cardiac arrest state, refusal to provide consent to participate, and ST-segment elevation myocardial infarction (STEMI) detected in the initial ECG test were all exclusion criteria. After screening, all patients were randomly separated into the intervention and control groups. Both groups had to get ECGs every 15 minutes starting at baseline. The SP-ECG group used the device's time setting, while the control group did not. It was the doctor's order after the patient was allotted a bed (Figure 1).

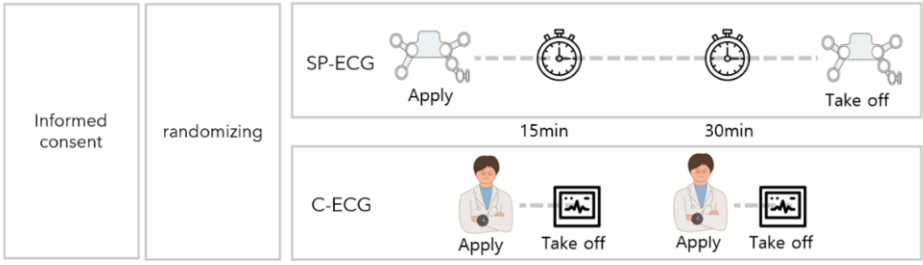


Figure 1. Construction of the study

Patients were randomly divided into two groups after screening. Both groups were instructed to perform ECGs 15 minutes and 30 minutes after the first ECG. The C-ECG group's patients used standard ECG equipment, while the SP-ECG patients had their ECGs recorded automatically using the device's timer. The SP-ECG was authorized as a Holter ECG by the Korean Ministry of Food and Drug Administration (Figure 2). The main body fits into the patch's socket. The tablet and main body are connected through Bluetooth to perform 12-lead ECG exams. The device's main board is a context-m4 DSP board. Analog front-end receives ECG signal from patch via analog-to-digital converter.[4] By clicking the tablet's “upload” button, researchers can transfer ECG exam results in real-time to their dashboards through LTE networks [5].

Follow-up timeliness When an ECG is taken within 3 minutes of the predetermined time. An ECG recorded 14 minutes after the initial ECG is considered timely. Based on initial triage information, age, gender, KTAS, heart rate, body temperature, respiration rate, and blood pressure were recorded.

A paired t-test was used to compare each patient's scores. P .05. The two groups were compared using two-way repeated-measures analysis. Neither group showed statistically significant differences over time, hence no post-analysis was undertaken.

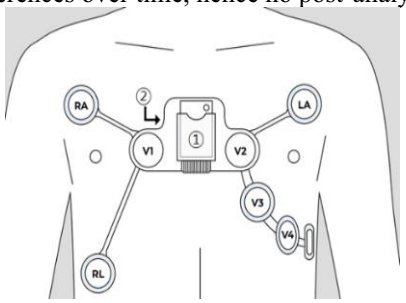


Figure 2. Configuration of the device

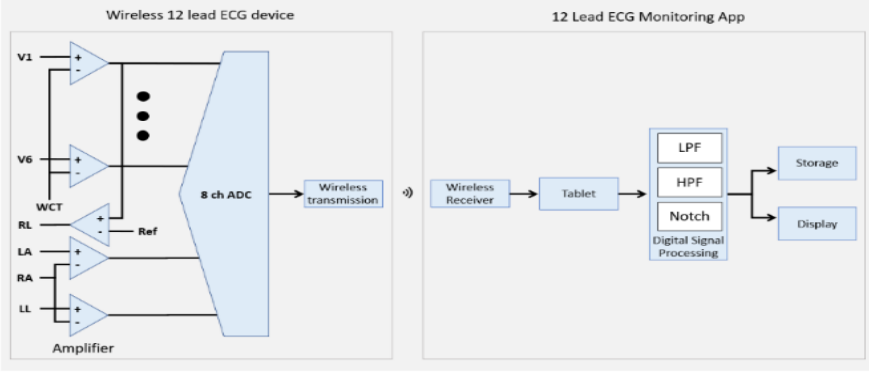


Figure 3. System architecture of Simple patch 12 lead ECG

Amplifier: Amplify analog voltage obtained from Electrode, 8ch ADC(8-channel analog-to-digital converter): Converting an amplified analog signal into a digital signal, Wireless transmission: Wireless transmission of converted digital signals, Wireless Receiver: Receive digital signals wirelessly, Digital signal processing: Compute received digital signals as ECG signals through digital operation, LPF (Low pass filter): Eliminates high frequency noise, HPF (High pass filter) : Eliminates low frequency noise, Notch (Notch filter): Eliminate noise at a certain frequency, eliminate 60 Hz noise used for commercial power sources, Storage: Store processed ECG data, Display: Output processed ECG data into a visualization graph

3. Results

The study included 36 people. The SP-ECG and C-ECG groups had median ages of 63.8 and 58.8 years. The study includes 33 patients out of 36 enrolled. One of the excluded patients requested out due to disorientation. A research breach caused time measurement problems in one person from each group. The final 33 participants averaged 61.06 years (SD: 15.8 years). 14 (42.4%) patients were female, with a KTAS of 3. Other traits showed no significant intergroup variations (Table 1).

Table 1. Demographic information of study participants

	SP-ECG group (N = 17)	C-ECG group (N = 16)	P-value
Age (Yr), mean ± SD	63.7 ± 18.4	58.1 ± 12.5	0.32
Gender			
Female, N (%)	6 (35.2)	8 (50.0)	
Male, N (%)	11 (64.8)	8 (50.0)	
KTAS, N (%)			0.80
-1	0	0	
-2	3 (17.6)	6 (37.5)	
-3	12 (70.6)	5 (31.2)	
-4	2 (11.8)	5 (31.2)	
-5	0	0	
Heart rate (bpm ± SD)	79.4 ± 17.9	79.3 ± 12.6	0.99
Body temperature (°C ± SD)	36.7 ± 0.5	36.7 ± 0.4	0.82
Respiratory rate (Breath/min ± SD)	18.5 ± 2.7	18.1 ± 1.7	0.61
SBP (mmHg ± SD)	133.4 ± 18.4	129.8 ± 18.2	0.57
DBP (mmHg ± SD)	79.7 ± 14.3	81.3± 14.6	0.75

a KTAS : Korean Triage and Acuity Scale b SBP : systolic blood pressure c DBP: diastolic blood pressure

3.1. Main Outcome

Figure 4 compares the time of the two research groups. The first follow-up ECG recording was timely in 87.5 percent of SP-ECG and 47.0 percent of C-ECG groups (P.74). At the second follow-up, it was 75.0% and 35.2% (P.71). Overall, 81.2 percent and 41.1 percent accuracy (P.62). The chart also shows that the C-ECG group's timing was not only outside the specified time window but was also severely delayed. Notably, four C-ECG subjects (4/16) had ECGs recorded more than an hour. Despite of the results, the minor issues have been detected with SP-accuracy. ECG's The device setup and study had four human errors in the test group. The provider should be adequately trained to avoid possible problems during the application. Later, more correction methods will be required. But more research is needed [4].

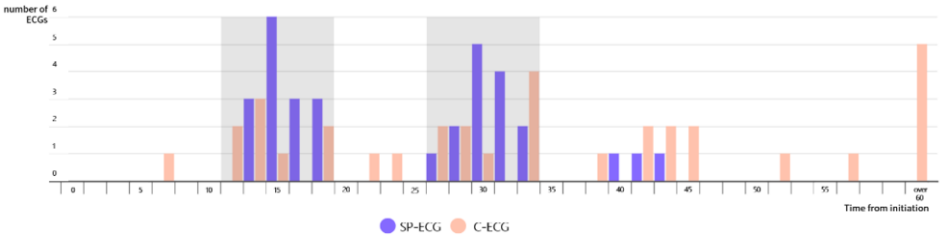


Figure 4. Patients' numbers on time accuracy
SP-ECG: Single Patch 12lead Electrocardiogram C-ECG: Conventional Electrocardiogram

4. Discussion

In addition to identifying substantial results and condensing the difficult process in the emergency room, this study is the first RCT using wireless and single patch ECG in Korea. The device could perform other procedures such as chest X-rays and lab testing without disrupting the emergency department's protocol. No patients had to leave the

study to follow the protocols, and no patients or medical staff had to miss other appointments.
Its correctness has yet to be verified, and small issues were discovered throughout the test. The supplier will need to be properly trained and compensated for future blunders. Human mistake occurred in setting up equipment and conducting research among four subjects in the test group, and to prevent future errors, the provider should be well trained before usage. Human resources will be needed to detect Single patch wireless 12lead ECG [6].

5. Limitations

The study has certain limitations. To begin, the study's patient cohort is not demographically representative. So, we should cross-validate our data with other institutions before drawing broad conclusions. Second, the trial was short-lived, therefore the procedure's impact in the ED couldn't be properly assessed. Finally, just time accuracy was compared and evaluated, leaving out the impact on the test provider's satisfaction, patient outcome, or diagnosis.

6. Conclusions

The ECG is the most significant and commonly done test in the ED to diagnose chest discomfort. The EKG's ST elevation helps distinguish STEMI patients from non-STEMI patients who require rapid treatment. However, about 40% of STEMI patients do not have ST elevation soon after symptom start. So, not only once, but frequently [7]. However, due to ED congestion and staffing shortages, some patients may not receive ECG results on time. To increase ECG timeliness, medical devices must be developed and used to autonomously capture patient ECGs without disrupting ED activities. The patient's examination and treatment processes were significantly faster with minimal disruption.

References

[1] Maršánová L, et al. ECG features and methods for automatic classification of ventricular premature and ischemic heartbeats: A comprehensive experimental study. Scientific reports. 2017 Sep 11;7(1):1-1.
[2] Antman EM, et al. ACC/AHA guidelines for the management of patients with ST-elevation myocardial infarction: a report of the American College of Cardiology/American Heart Association Task Force on Practice Guidelines (Committee to Revise the 1999 Guidelines for the Management of Patients with Acute Myocardial Infarction). Journal of the American college of cardiology. 2004 Aug 4;44(3):E1-211.
[3] Hong KJ, et al. Association between ED crowding and delay in resuscitation effort. Am J Emerg Med. 2013 Mar;31(3):509-15. doi: 10.1016/j.ajem.2012.09.029. Epub 2012 Nov 15. PMID: 23159432.
[4] Nemati E, Deen MJ, Mondal T. A wireless wearable ECG sensor for long-term applications. IEEE Communications Magazine. 2012 Jan 5;50(1):36-43.
[5] Mankodiya K, et al. Wearable ECG module for long-term recordings using a smartphone processor. In Proceedings of the 5th International Workshop on Ubiquitous Health and Wellness, Copenhagen, Denmark 2010 Sep (Vol. 2629).
[6] Ouellette L, et al. Emergency department electrocardiogram images sent through the mobile phone: Feasibility and accuracy. The American journal of emergency medicine. 2018 Apr;36(4):731-2.
[7] Fakhri Y, et al. Evaluation of acute ischemia in pre-procedure ECG predicts myocardial salvage after primary PCI in STEMI patients with symptoms >12hours. J Electrocardiol. 2016 May-Jun;49(3):278-83.

Challenges of Trustable AI and Added-Value on Health
B. Séroussi et al. (Eds.)

doi:10.3233/SHTI220628

Assessing the Quality of Direct-to-Consumer Teleconsultation Services in Canada

Jean Noel NIKIEMA[a,1], Eleah STRINGER[b], Marie-Pierre MOREAULT[a], Priscille PANA[a], Marco LAVERDIERE[c], Catherine RÉGIS[d], Jean-Louis DENIS[a], Béatrice GODARD[a], Mylaine BRETON[e], Guy PARÉ[f], Aviv SHACHAK[g], Claudia LAI[g], Elizabeth M. BORYCKI[b], Andre W. KUSHNIRUK[b] and Aude MOTULSKY[a]

[a] *School of Public Health, Université de Montréal, Canada*
[b] *School of Health Information Science, University of Victoria, Canada*
[c] *Faculty of Law, Université de Sherbrooke, Canada*
[d] *Faculty of Law, Université de Montréal, Canada*
[e] *Faculty of Medicine and Health Sciences, Université de Sherbrooke, Canada*
[f] *HEC Montréal, Canada*
[g] *Institute of Health Policy, Management, and Evaluation, University of Toronto, Canada*

Abstract. The objective of this study was to describe and assess the quality of the direct-to-consumer medical teleconsultation landscape in three Canadian provinces. An environmental scan of primary care teleconsultation platforms was conducted in January 2022 to identify medical teleconsultation platforms in Quebec (Qc), Ontario, and British Columbia (BC). The quality of each teleconsultation platform was assessed using a modified version of the HONcode principles. Nineteen different direct-to-consumer medical teleconsultation platforms were identified across the three provinces. The quality of these teleconsultation platforms was very heterogeneous. The landscape of virtual primary care is changing rapidly in the Canadian ecosystem, and the transparency of current teleconsultation platforms could be improved.

Keywords. Teleconsultation, Honcode, primary care, virtual care.

1. Introduction and Methods

The COVID-19 pandemic has accelerated the transformation of medical consultation towards virtual care to reduce physical contacts [1]. In Canada, this meant accessing physicians primarily by telephone [2], while the transformation via online tools was driven by private enterprise [3]. With a universal but narrow health care system, (a limited number of health care services are universally accessible), the organization of the Canadian health care services is widely heterogeneous across provincial and territorial jurisdictions, especially for primary care access to medical consultation. While the rapid emergence of privately driven virtual clinics offered new ways of accessing a general

[1] Corresponding Author, DGEPS, Université de Montréal; E-mail: jean.nikiema@umontreal.ca
The work was made possible by funding from OBVIA and the Fonds de recherche du Québec (FRQ)

practitioner, these novel channels of accessing medical expertise introduce unique challenges to the universal health care system. The objective of this study was to describe and assess the quality of the changing direct-to-consumer medical teleconsultation landscape across three Canadian provinces - Quebec (Qc), Ontario and British Columbia (BC) - in 2022. An environmental scan of primary care teleconsultation platforms in each participating province was conducted. A list of 19 different direct-to-consumer medical teleconsultation platforms were identified and assessed for quality using an adapted version of the Health on the Net Code (HONcode) principles (labtns.ca/HONcode/).

2. Results and Discussion

Only four teleconsultation platforms were in a Business-to-Business (B2B) model (for private insurance and/or employers). While most of them were only virtual, some were associated with physical clinics who provide virtual care in addition to their standard in-person consultations. Some teleconsultation platforms were covered by provincial public health insurance, most of them were accessible through direct patient contribution (ranging from $20 to $120 per consultation), either on a fee-for-service basis, or through monthly or annual subscriptions. This illustrates that, for the moment, it is unclear how teleconsultations fit within Canada's universal healthcare system, where medical consults are considered an 'essential' service that should usually not be restricted by fee for access. As the nature of teleconsultation platforms varies by provincial jurisdiction, business models were difficult to describe. Further research is needed to analyze the impact of these differences by province on equity in access to primary care. The compliance with the modified HONcode principles was very heterogeneous. It was particularly difficult sourcing details about participating practitioners, with only 31% (n=6) of teleconsultation platforms detailing the qualifications and location of their practitioners. While several teleconsultation platforms provided names, photographs, diplomas, etc. for their practitioners, other offered a general, blanket statement on their practitioners being authorized to provide medical services in Canada. When the backbone of Canadian primary care is based on an established relationship between family physicians (or nurse practitioners) and their patients, the complementarity of teleconsultation platforms is of utmost importance. While no formal integrative mechanism was discussed by any teleconsultation platform, 47% (n=9) clearly delineated the boundary of their service and interactions with primary providers (e.g. a patient's family doctor) contributing to continuity of care by providing a clear recommendation for services not offered. Just over half of the teleconsultation platforms (n=11) provided recommendations for emergency scenarios which, we argue, are required for patient safety (e.g. "call 911 in case of emergency"). Policy and regulation should be prioritized to ensure safe and equitable access to medical primary care.

References

[1] Hollander JE, Carr BG. Virtually perfect? Telemedicine for COVID-19. New England Journal of Medicine. 2020;382(18):1679-81.
[2] Breton M, et al. Telehealth in Primary Healthcare: A Portrait of its Rapid Implementation during the COVID-19 Pandemic. Healthcare Policy. 2021;17(1):73.
[3] Matthewman S, et al. An Environmental Scan of Virtual "Walk-In" Clinics in Canada: Comparative Study. Journal of Medical Internet Research. 2021;23(6):e27259.

Challenges of Trustable AI and Added-Value on Health
B. Séroussi et al. (Eds.)

doi:10.3233/SHTI220629

SPEAKapp: Language as Digital Biomarker for Mental Health

Chiara BARATTIERI DI SAN PIETRO[a,1], Valentina SIMONETTI[a] and Maria BULGHERONI[a]
[a] *Ab.Acus s.r.l.m Milan, Italy*

Abstract. Language has always been a rich source of amnestic information in the context of cognitive and affective disorders. Many of the language tests and metrics that are employed today have been designed when the only available technology was paper and pencil. Here we present a novel digital tool to administer and score a set of tests based on language.

Keywords. Mental Health, Natural Language Processing, mHealth

1. Introduction

Language dysfunctions are common features in many psychopathologies. Part of the reason why is that language requires the simultaneous activation of multiple cognitive systems [1]. The assessment of language is generally carried out relying on the subjective appreciation by the clinician.

Recent advances of Natural Language Processing (NLP) techniques, Voice Analysis, and Automatic Speech Recognition have gained the attention of clinicians working with people with cognitive and affective disturbances [2]. Unfortunately, the practical adoption of the solutions proposed into clinical setting is limited by the need of specific recording equipment and the availability of trained staff. To overcome the barrier between technological advancement and clinical needs, we developed SPEAKapp, a system based on a mobile application to deliver and analyze speech and language data for clinical and research purposes.

2. Methods

SPEAKapp implements five tests: verbal fluency, automatic series, repetition of complex sentences, picture naming, and prose recall. Each test has been chosen to cover a broad range of cognitive functions and systems affected in different psychiatric and neurological conditions. The system collects this verbal production, and it is the starting point of the features extraction process. The text content of each audio is first extracted using a Speech-To-Text module that leverages commercial Google Speech-To-Text APIs (https://cloud.google.com/speech-to-text). Then, both the raw audio and the text

[1] Corresponding Author, Chiara Barattieri di San Pietro, Ab.Acus s.r.l., via Caracciolo 77, 20100 Milan, Italy, chiarabarattieri@ab-acus.eu.

content are processed to extract acoustic and semantic features. The system's module for the extraction of voice's acoustic signal parameters computes acoustic features commonly used in acoustic analysis applications: fundamental frequency, jitter, shimmer, harmonicity to noise ratio, and phonation duration. The SPEAKapp acoustic analysis stands upon the Open Source Parselmouth Python library for Praat [3], while novel NLP-based scores are based on Distributional Semantic Models of language. The DSMs used in SPEAKapp were created with word2vec [4]. Features extracted by SPEAKapp uses FHIR (Fast Health Interoperability Resources) standard and can be directly integrated with any Electronic Health Record system supporting this standard.

3. Expected Results

Scoring to the neuropsychological tests collected via the mobile phone app will be calculated twice: i) manually, after having manually transcribed the answers, following the gold standard in the literature and ii) via the app software: the system implements the logics needed to compute the scoring as per literature, but based on the automatic transcription. To test the convergent validity of the system, the correlation between the scoring computed manually and the result of the automatic elaboration will be calculated, according to the Cronbach's α.

4. Discussion and Conclusions

Natural language use can be a sensitive and nonintrusive indicator of changes in internal affective states such as stress, depression, or anxiety. The integration of novel indexes based on verbal performance with standardized measures might lead to novel insights into mental health conditions, as well as to the identification of light and reliable indexes of cognitive functioning. In the long term, identifying a valid marker of treatment efficacy would support clinical research and innovation, facilitating the evaluation of new drugs' efficacy and the mental health of the population served.

Acknowledgments

SPEAKapp has received funding from the European Union s Horizon 2020 research and innovation programme under grant agreement Nº 857223.

References

[1] Friederici AD. The cortical language circuit: from auditory perception to sentence comprehension. Trends in cognitive sciences. 2012 May 1;16(5):262-8..
[2] Cummins N, Scherer S, Krajewski J, Schnieder S, Epps J, Quatieri TF. A review of depression and suicide risk assessment using speech analysis. Speech communication. 2015 Jul 1;71:10-49..
[3] Boersma P. Praat: doing phonetics by computer [Computer program]. http://www. praat. org/. 2011.
[4] Mikolov T, Chen K, Corrado G, Dean J. Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781. 2013 Jan 16.

Challenges of Trustable AI and Added-Value on Health
B. Séroussi et al. (Eds.)

doi:10.3233/SHTI220630

Pervasive Monitoring of Public Health and Well-Being in Urban Areas with Blue-Green Solutions

Parisis GALLOS[a,b,1], Andreas MENYCHTAS[a,b], Christos PANAGOPOULOS[a], Maria KASELIMI[c], Ioannis RALLIS[c], Anastasios DOULAMIS[c], Nikolaos DOULAMIS[c], Manthos BIMPAS[c], Aikaterini AGGELI[d], Eftychios PROTOPAPADAKIS[c], Emmanuel SARDIS[c] and Ilias MAGLOGIANNIS[b]
[a] *BioAssist S.A., Athens, Greece*
[b] *Computational Biomedicine Research Lab, Department of Digital Systems, University of Piraeus, Piraeus, Greece*
[c] *National Technical University of Athens, Athens, Greece*
[d] *National and Kapodistrian University of Athens, Greece*

Abstract. The urban environment seems to affect the citizens' health. The implementation of Blue-Green Solutions (BGS) in urban areas have been used to promote public health and citizens well-being. The aim of this paper is to present the development of an mHealth app for monitoring patients and citizens health status in areas where BGS will be applied. The "HEART by BioAsssist" application could be used as a health and other data collection tool as well as an "intelligent assistant" to monitor and promote patient's physical activity in areas with Blue-Green Solutions.

Keywords. Blue-Green Solutions, Urban Health, Well-Being, Pervasive Computing

1. Introduction

The urban environment seems to affect the citizens' health. The implementation of Blue-Green Solutions (BGS) in urban areas have been used to promote public health and citizens well-being [1]. The mobile health (mHealth) technologies contribution in people's well-being has been examined and found to be significant based on the international literature [2]. The aim of this paper is to present the development of an mHealth app for monitoring patients and citizens health status in areas where BGS will be applied in the context of the HEART project.

2. Methods

The application requirements were to monitor users' physical, physiological, and emotional status using wearable devices, as well as to motivate users to engage in social

[1] Corresponding Author, Parisis Gallos; E-mail: parisgallos@bioassist.gr.

activities. The specifications of the application include the recording of heartrate, SpO_2, sleep quality, stress levels and physical activity including the number of steps, daily exercise, walking/running and others.

3. Results and Discussion

The "HEART by BioAsssist" application was developed to monitor the users' daily physical activity and other variables related to their health status in areas where BGS will be applied. The application is available for multiple smartphone devices, and it is compatible with most commercially available wearables. The application development was based on an integrated platform approach was used in previous studies [3,4]. HL7-FAIR, standards have been adopted for data storage and data exchange. The "HEART by BioAsssist" application could be used as a health and other data collection tool as well as an "intelligent assistant" to monitor and promote patient's physical activity. The proposed application may be also used to support clinical studies requirements for remote monitoring, advanced study management, and higher patient adherence and compliance. Relative studies have used comparable approaches to monitor people's well-being and public health using similar mHealth applications [2,5].

4. Conclusions

The aforementioned application is developed in the frame of the EU-funded project "HEAlthier Cities through Blue-Green Regenerative Technologies: the HEART Approach" as health and personal data tool. HEART's integrated approach aims to systematically improve urban health using BGS on future area planning.

Acknowledgements

This work is co-funded by the Horizon 2020 Programme of the European Commission-EU.3.1.2 Grant Agreement number 945105 - HEAlthier Cities through Blue-Green Regenerative Technologies: the HEART Approach.

References

[1] van den Bosch M, Ode Sang Å. Urban natural environments as nature-based solutions for improved public health - A systematic review of reviews. Environ Res. 2017 Oct;158:373-384.
[2] Saxena M, Saxena A. Evolution of mHealth eco-system: a step towards personalized medicine. Adv Intell Syst Comput. 2020; 1087: 351-370.
[3] Panagopoulos C, Malli F, Menychtas A, Smyrli EP, Georgountzou A, Daniil Z, Gourgoulianis KI, Tsanakas P, Maglogiannis I. Utilizing a Homecare Platform for Remote Monitoring of Patients with Idiopathic Pulmonary Fibrosis. Adv Exp Med Biol. 2017;989:177-187.
[4] Gallos P, et al. Quantifying Citizens' Well-Being in Areas with Natural Based Solutions Using Mobile Computing. Stud Health Technol Inform. 2022 Jan 14;289:465-468.
[5] Freudenberg N. Assessing the Public Health Impact of the mHealth App Business. Am J Public Health. 2017 Nov;107(11):1694-1696.

Challenges of Trustable AI and Added-Value on Health
B. Séroussi et al. (Eds.)

doi:10.3233/SHTI220631

Classification of Dichotomous Emotional States Using Electrodermal Activity Signals and Multispectral Analysis

Yedukondala Rao VEERANKI[a,1], Nagarajan GANAPATHY[b] and Ramakrishnan SWAMINATHAN[a]
[a] *Biomedical Engineering Group, Department of Applied Mechanics, Indian Institute of Technology Madras, Chennai, Tamil Nadu, India*
[b] *Peter L. Reichertz Institute for Medical Informatics of TU Braunschweig and Hannover Medical School, Germany*

Abstract. In this work, an analysis based on complex demodulation is proposed to classify dichotomous emotional states using Electrodermal activity (EDA) signals. For this, annotated happy and sad EDA is obtained from an online public database. The sympathetic activity indices, namely Time-varying (TVSymp) and Modified TVSymp, are computed from the reconstructed EDA signal. Further, the derivative of phasic EDA is calculated from the phasic component obtained using the convex optimization (cvxEDA) based EDA decomposition method. Five statistical features are computed from each index and used for the classification. The results of the classification indicate that these features are capable of differentiating happy and sad emotional states with 75% accuracy. This technique could be effective in the identification of clinical disorders associated with happy and sad emotional states.

Keywords. Electrodermal activity, multispectral analysis, support vector machine

1. Introduction

Emotion is a conscious phenomenon that plays an indispensable role in social interaction, quality of life, and decision-making. Electrodermal activity (EDA) based emotion detection is the most popular due to its simple instrumentation and non-intrusiveness. This work explores the feasibility of variable frequency complex demodulation (VFCDM) [1] based analysis to classify dichotomous emotional states in EDA signals.

2. Methodology

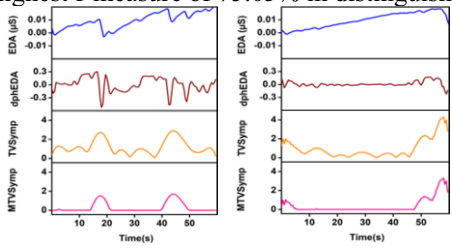
For this study, happy and sad EDA are collected from the DEAP database [2] that are acquired from 32 healthy volunteers while watching a 60s long happy and sad video separately. The EDA is downsampled to 2Hz and then high pass filtered with a cut-off

[1] Corresponding Author, Yedukondala Rao Veeranki; Biomedical Engineering Group, Dept. of Applied Mechanics, Indian Institute of Technology Madras, Chennai, India; E-mail: ykraoveeranki@gmail.com.

frequency of 0.01Hz. VFCDM is used to reconstruct EDA signals with frequencies ranging from 0.045 to 0.25 Hz. The TVSymp or instantaneous amplitude of the reconstructed signal is obtained using the Hilbert transform. MTVSymp is computed by subtracting the instantaneous amplitude value with respect to its mean value [1]. Further, the EDA is decomposed into the phasic component using the cvxEDA [3] method, and its derivative (dphEDA) is calculated using the 5-point stencil equation [1]. Five statistical features, namely standard deviation, skewness, kurtosis, mean, and RMS, are computed from the estimated SNS activity indices. The features extracted from each index are used for the classification. The performance of the support vector machine (SVM) classifier is evaluated with a k-fold cross-validation procedure with k=10.

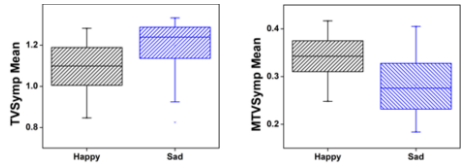
3. Results and Discussion

Figures 1(a) and 1(b) illustrate the representative dphEDA, TVSymp, and MTVSymp indices computed from the EDA signal for happy and sad emotional states, respectively. The EDA of happy stimulus-response has higher variations than sad stimulus-response. The SNS activity indices capture the fluctuations in the EDA signals. It is found that the amplitude and fluctuations of dphEDA index during happiness are higher compared to sadness. TVSymp detects the envelope of the reconstructed EDA signal. It is also seen that the MTVSymp produces non-negative output. The boxplot representation of the mean feature obtained TVSymp and MTVSymp are shown in Fig. 2(a) and Fig. 2(b), respectively. From the box plot of TVSymp Mean, it is seen that the mean and median values of the sad class are higher than happy. From the box plot of MTVSymp Mean, It is found that the mean and median of the happy class are high. The SVM yields the highest f-measure of 75.05% in distinguishing happy and sad emotional states.



(a) (b)

Figure 1. Representative EDA signal and the corresponding dphEDA, TVSymp, and MTVSymp for (a) happy and (b) sad emotions.



(a) (b)

Figure 2. Boxplot representation of the mean of (a) TVSymp, and (b) MTVSymp in differentiating dichotomous emotional states.

References

[1] Kong Y, Posada-Quintero HF, Chon KH. Real-Time High-Level Acute Pain Detection Using a Smartphone and a Wrist-Worn Electrodermal Activity Sensor. Sensors. 2021 Jan;21(12):3956.

[2] Koelstra S, Muhl C, Soleymani M, Lee JS, Yazdani A, Ebrahimi T, Pun T, Nijholt A, Patras I. DEAP: A database for emotion analysis using physiological signals. IEEE transactions on affective computing. 2011 Jun 9;3(1):18-31.

[3] Greco A, Valenza G, Lanata A, Scilingo EP, Citi L. cvxEDA: A convex optimization approach to electrodermal activity processing. IEEE Transactions on Biomedical Engineering. 2015 Aug 28;63(4):797-804.

Challenges of Trustable AI and Added-Value on Health
B. Séroussi et al. (Eds.)

doi:10.3233/SHTI220632

Classification of Emotional States Using EEG Signals and Wavelet Packet Transform Features

Himanshu KUMAR[a,1], Nagarajan GANAPATHY[b], Subha D PUTHANKATTIL[c] and Ramakrishnan SWAMINATHAN[a]
[a] *Biomedical Engineering Group, Department of Applied Mechanics, Indian Institute of Technology Madras, Chennai, India*
[b] *Peter L. Reichertz Institute for Medical Informatics of TU Braunschweig and Hannover Medical School, Germany*
[c] *Department of Electrical Engineering, National Institute of Technology Calicut, Kozhikode, India*

Abstract. In this work, an attempt has been made to classify arousal and valence states of emotion using time-domain features extracted from the Wavelet Packet Transform. For this, Electroencephalogram (EEG) signals from the publicly available DEAP database are considered. EEG signals are first decomposed using wavelet packet decomposition into θ, α, β, and γ bands. Then featural, namely band energy, sub-band energy ratio, root mean of energy, and information entropy of band energy is estimated. These features are fed into various machine learning classifiers such as support vector machines, linear discriminant analysis, K-nearest neighbor, and random forest. Results indicate that features extracted from wavelet packet transform can predict the arousal and valence emotional states. It is also seen that Support Vector Machines perform the best for both arousal (f-m = 75.68%) and valence(f-m=57.53%). This method can be used for the recognition of emotional states for various clinical purposes in emotion-related psychological disorders like major depressive disorder.

Keywords. Emotion, Electroencephalogram, Wavelet Packet Transform

1. Introduction

Emotion reflects human feelings, thoughts, and behavior [1]. There are two emotion classification models, including basic emotion-based and dimensional space-based emotion classification. The 2-D spatial model combined with valence and arousal is the most used in the dimension space model [1]. Traditional Electroencephalogram (EEG) based emotion recognition algorithms mainly extract single EEG features from the time domain, frequency domain, or information entropy, which cannot achieve satisfactory performance [2]. Recently wavelet transformed-based features have been used to classify emotion to achieve better performance [2].

[1] Corresponding Author, Himanshu Kumar, Biomedical Engineering Group, Department of Applied Mechanics, Indian Institute of Technology Madras, Chennai 600036, India; E-mail: him241994@gmail.com.

2. Materials and Methods

EEG signals from the DEAP database are considered for this study [3]. Signals from 32 subjects recorded over 32 electrodes for 60 seconds, sampled at 128 Hz, were used. Subjects annotated signals as low or high arousal and low or high valence using a self-assessment manikin. Signals from all electrodes and all 32 subjects are subjected to Wavelet Packet Transform to extract the EEG frequency band, including θ wave, α wave, β wave, and γ wave. Wavelet coefficients are used to calculate the band energy, the sub-band energy ratio, root mean of energy, and EEG energy entropy. The features are fed into four classifiers: support vector machines (SVM), linear discriminant analysis (LDA), K-nearest neighbor (KNN), and random forest (RF). Feature selection algorithm, namely support vector machine – recursive feature elimination (SVM-RFE), is used to reduce the features [4]. The classifiers are cross-validated using the k-fold cross-validation technique (k=10). Accuracy (A), Precision (P), Recall (R), and F-measure (f-m) are used to evaluate the classifiers.

3. Results and Discussion

Table 1 shows the classification performance for classifying arousal and valence emotional states. Results indicate that SVM outperforms all other classifiers. In the case of valence, SVM-RFE achieved a similar performance with the top 100 selected features and arousal with similar performance with the top 10 features only.

Table 1. Classification performance by the various classifier for classifying arousal and valence emotional states

	Valence				Arousal			
	P	**R**	**A**	**f-m**	**P**	**R**	**A**	**f-m**
RF	57.23	46.87	59.30	51.53	68.74	76.49	63.20	72.41
LDA	52.86	50.08	56.33	51.43	69.53	70.05	61.72	69.79
SVM	**58.23**	**56.85**	**61.25**	**57.53**	66.82	**87.25**	**64.61**	**75.68**
KNN	54.21	52.28	57.58	53.23	**68.92**	77.10	63.59	72.78
SVM-RFE	58.60	58.21	61.72	58.40	66.48	88.37	64.53	75.88

4. Conclusions

The proposed approach is able to classify the arousal and valence states of emotion using Wavelet Packet Transform features and can be helpful in behavioral analysis.

References

[1] Posner J, Russell JA, Peterson BS. The circumplex model of affect: An integrative approach to affective neuroscience, cognitive development, and psychopathology. Development and psychopathology. 2005;17(3):715.

[2] Liu Y, & Fu G. Emotion recognition by deeply learned multi-channel textual and EEG features. Future Gen Comp Sys 2021;119:1–6.

[3] Koelstra S, Muhl C, Soleymani M, Jong-Seok Lee, Yazdani A, Ebrahimi T, et al. DEAP: A Database for Emotion Analysis ;Using Physiological Signals. IEEE Trans Affective Comput. 2012 Jan;3(1):18–31.

[4] Sanz H, Valim C, Vegas E, Oller JM, Reverter F. SVM-RFE: selection and visualization of the most relevant features through non-linear kernels. BMC Bioinformatics. 2018 Nov 19;19(1):432.

Challenges of Trustable AI and Added-Value on Health
B. Séroussi et al. (Eds.)

doi:10.3233/SHTI220633

Project iFoot - Optimising the Treatment of Patients with Diabetic Foot Syndrome with an Intelligent Dressing

Dominik LUDMANN[a,1], Timo TROMP[a] and Hubert OTTEN[a]
[a]*Hochschule Niederrhein, University of Applied Sciences, Faculty of Healthcare, Germany*

Keywords. diabetic foot syndrome, sensorics, telemedicine, E-Health

1. Introduction

Diabetes mellitus is one of the world's most serious noncommunicable, chronical diseases and poses a serious public health challenge [1]. An estimated number of 8.5 million Germans currently live with diabetes Type 2 and projections indicate that by 2040, these numbers might increase up to 12.3 million cases [2]. The diabetic foot syndrome (DFS) is a frequent complication of diabetes mellitus and is a collective term for various pathological changes of the foot of people with diabetes mellitus. About 250,000 people in Germany develop a diabetic foot ulcer (DFU) every year, resulting in 13,000 major amputations [3]. The costs of treating diabetic foot syndrome in Germany amount to approximately 2.5 billion euros per year and therefore are of great importance for the health care system [4]. The iFoot research project focuses on an optimized approach to DFS care and has therefore developed an innovative system solution that aims to improve the treatment of DFS by measuring medically relevant values in wound area.

2. Methods

In a prospective, randomized study from February to October 2021 twenty participants between the age of 18 and 85 with a diagnosed form of DFS were included. The primary outcome was time until the wound area was halved; the wound size was documented every 14 days during control visits. Study participants were equipped with a sensor-based intelligent dressing and a commercially available smartwatch. Sensor measurements of pressure, temperature and humidity were recorded near the wound every ten seconds over the entire duration of the study and complemented by the participant's step counts. Participants of the intervention group received audio-visual warning messages via a smartwatch-application in case of exceeding a predefined pressure limit. The aim of this information given to the participants, was to provide a biofeedback in order to

[1] Corresponding Author, Dominik Ludmann, Hochschule Niederrhein – University of Applied Sciences Krefeld, Reinarzstraße 49, 47805 Krefeld, Germany; E-mail: dominik.ludmann@hsnr.de.

compensate for the given incomplete sensation in the feet of participant's due to diabetes-related polyneuropathies.

3. Results and Discussion

Participants of the intervention group were able to adjust their behavior by responding to the biofeedback delivered by the smartwatch, enabling them to avoid incorrect pressure load on their feet and resulting in a potentially positive influence on the healing process of the wound. An example is given in figure 1 which shows the pressure measurements of one participant who, after approximately one week during which numerous alarms were triggered, began to consistently reduce the pressure load on his wound. As a result, the circumstances for the healing process of the existing wound seem to have improved and a swift wound closure was detected.

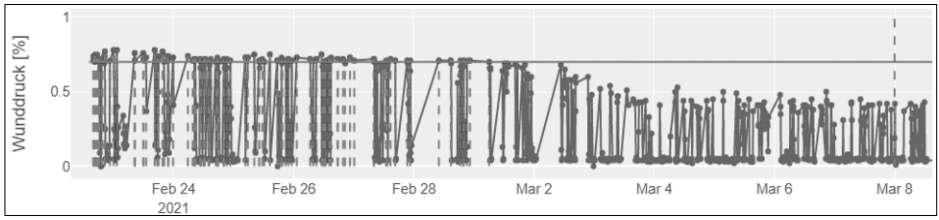


Figure 1. Pressure measurement with alarms shown as dashed lines

Several technical concepts to measure the pressure load on the feet of those who suffer from DFS are published, some are commercially available and mainly serve the purpose of preventing diabetic foot ulcers [5][6]. In distinction to this, the iFoot-system was developed to provide a sensor-based approach to treat DFS with active patient participation.

4. Conclusions

The case-by-case analysis of the recently completed pilot study shows that the use of the iFoot-system is suitable for sensor-based treatment of patients with DFS. In addition to individual case reports, further data evaluations are pending.

References

[1] International Diabetes Federation. IDF Diabetes Atlas, 10th edn. Brussels, Belgium: 2021. Available at: https://www.diabetesatlas.org [accessed january 21, 2022]
[2] Tönnies T, Röckl S, Hoyer A, Heidemann C, Baumert J, Du Y, et al. Projected number of people with diagnosed type 2 diabetes in Germany in 2040. Diabetic Medicine. 2019;36(10):1217–25.
[3] Morbach S et al., Diabetic Foot Syndrome - Part 1 Definition, pathophysiology, diagnostics and classification, Springer Switzerland, 2020.
[4] Hochlenert D, et al. Diabetic Foot Syndrome: From Entity to Therapy. Cham: Springer, 2018.
[5] Raviglione, A., Reif, R., Macagno, M. et al. (2017). Real-time smart textile-based system to monitor pressure offloading of diabetic foot ulcers. J. Diabetes Sci. Technol. 11 (5):894–898.
[6] Abbott CA, et al. Innovative intelligent insole system reduces diabetic foot ulcer recurrence at plantar sites: a prospective, randomised, proof-of-concept study. Lancet Digit Health. 2019 Oct;1(6):e308-e318.

Challenges of Trustable AI and Added-Value on Health
B. Séroussi et al. (Eds.)

doi:10.3233/SHTI220634

Electronic Health Records on the Top of Medical Device Incident Reports

Krista KUMPUVAARA[a,b,1], Vesa JORMANAINEN[c,d], Tarja VAINIOLA[e] and Tuija S. IKONEN[b,f]

[a] *Helsinki University Hospital, Hyvinkää, Finland*
[b] *University of Turku, Department of Public Health, Turku, Finland*
[c] *Finnish Institute for Health and Welfare, Helsinki, Finland*
[d] *University of Helsinki, Department of Public Health, Helsinki, Finland*
[e] *Finnish Medicines Agency Fimea, Medical Devices Unit, Helsinki, Finland*
[f] *Ostrobothnia Wellbeing Services County, Vaasa, Finland*

Abstract. Medical Device incident reporting is a legal obligation for professional users in Finland. We analyzed all medical device incident reports recorded into the national incident repository from January 2014 to August 2021. Almost 30% of the total of 5,897 recorded incidents were caused by top ten devices, of which electronic health records were the most common (332 incidents). High number of incidents caused by electronic health records arouses safety concerns. A further analysis is required to explore the causes of findings.

Keywords. Medical device, safety incident, electronic health records

1. Introduction

Submitting a medical device incident report in Finland is a legal obligation for professional users. Reporting an incident applies to all medical devices and requires health of a person being in danger and a problem in relation to a medical device, and situations that could possibly have endangered a person's health. By definition, Parts of the Electronic Health Records (EHR) belong to medical devices. In Finland, technology related patient safety incidents have been studied in hospital district's or regulatory authority's registers [1] but rarely comprising all medical device incidents [2].

In this study, we assess frequencies of top ten medical device incidents reported by professional users into the nationwide medical device incident database at Finnish Medicines Agency, the regulatory authority responsible for incident repository.

2. Methods

The incidence reports from professional users contain data on user and affiliation, the medical device concerned in detail, type and severity of harm and additional device

[1] Corresponding Author MD Krista Kumpuvaara, Affiliation, Sälpätie 11, 10700 Vantaa, Finland; E-mail: krista.t.kumpuvaara@utu.fi.

details (e.g., device maintenance, etc.). As part of the Prime Minister's Office development project [3], we analyzed professional users' medical device incident reporting data January 1, 2014 - August 10, 2021 in accordance to GDPR regulations.

3. Results

Altogether 5,897 medical device incident reports were recorded. Ten most often reported categories made up to 29.3% (1,725/5,897) of all incidents in the study period. Of the top ten incident categories, EHRs had the highest number of reports (n=332; 5.6%), followed by hip artificial joint (n=294; 5.0%) and patient bed (n=202; 3.4%) (Figure 1).

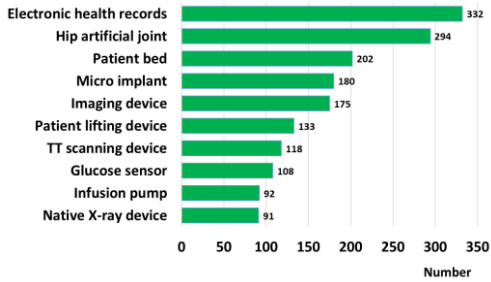


Figure 1. Ten most frequent medical device incident report categories recorded by professional users into the national incident repository from January 1, 2014 to August 10, 2021 (n=5,897).

4. Discussion and Conclusion

In Finland, the ten most reported devices were related to 30% of all incident reports, of which EHRs were the most frequent. Healthcare service providers use various EHR systems. Standardization of data structures, classifications and codes are inadequate. User interfaces are suboptimal for daily use, and systematic user education is a challenge [4]. Finland has introduced national, centralized, shared, integrated and interoperable electronic data system services for standardization and interoperability [5]. Further analyses are required to identify the root causes of reported medical device incidents.

References

[1] Palojoki S, Mäkelä M, Lehtonen L, Saranto K. An analysis of electronic health record–related patient safety incidents. Health Inform J 2017;23(2):134–145.

[2] Sievänen H, Pommelin P. Quality analysis of medical device vigilance reports. Technol Health Care 2003;11(4):275–281.

[3] Virkki M, Leskelä R-L, Ikonen T, Haatainen K, Welling M, Rauhala A, et al. Current situation of patient and client/customer safety and follow-up procedures in Finland: A suggestion for a measurement framework. Publications of the Government´s analysis, assessment and research activities 2021:68. Helsinki (Finland): Prime Minister's Office; 2021. http://urn.fi/URN:ISBN:978-952-383-334-0

[4] Kaipio J, Lääveri T, Hyppönen H, et al. Usability problems do not heal by themselves: National survey on physicians' experiences with EHRs in Finland. Int J Med Inform 2017;97:266-281.

[5] Jormanainen V. Large-scale implementation and adoption of the Finnish national Kanta services in 2010–2017: a prospective, longitudinal, indicator-based study. Finnish J eHealth eWelfare 2018;10(4):381–395.

Challenges of Trustable AI and Added-Value on Health
B. Séroussi et al. (Eds.)

doi:10.3233/SHTI220635

How to Publish Wearables' Data: Practical Guidelines to Protect User Privacy

Thomas MARCHIORO [a,b,1], Andrei KAZLOUSKI [a,b] and Evangelos MARKATOS [a,b]
[a] *Computer Science Department, University of Crete, Greece*
[b] *Institute of Computer Science, Foundation for Research and Technology*

Abstract. The spread of wearable fitness trackers has contributed to the increase of various fitness studies, utilizing such devices to collect data from a group of users. When these data are disclosed, the lack of sanitization can lead to severe privacy risks. In this paper, we discuss the various threats that are posed by disclosing unaltered information from wearable devices. We also dismiss common fallacies of fitness data sharing and present practical guidelines to preserve user privacy.

Keywords. Electronic Health Records, Data Privacy, Wearable Devices

1. Introduction

In lifelogging experiments the activity of a small group of people is monitored through wearable fitness trackers for a number of days. During this period, the participants provide a considerable amount of information, including surveys of microdata (e.g., age, gender, height, weight), as well as activity logs of steps, burned calories, heart rate, and more. When these data are aggregated and published, often little effort is put to really make the disclosed information "anonymous". As a result, details about the participants' activity and habits can potentially be leaked, compromising their privacy.

2. Privacy aspects of sharing data from wearables

Albeit well-known in privacy literature, sensitiveness of quasi-identifiers tends to be neglected by data publishers, likely with the intent of providing useful information about the dataset population. For example, most participants of prominent public life-logging datasets [1,2] can be de-anonymized based on quasi-identifiers, such as personal and physical characteristics. Applying standard anonymization to these attributes (e.g., using *k*-anonymity [3] or personalized privacy [4]) is thus a first step to protect users' privacy. However, this may not be enough to guarantee that a target is not found in the dataset. Activity logs themselves often contain information that is derived from an individual's attributes: calories, for instance, are often estimated by combining activity information with physical characteristics such as age, gender, height, and weight [5]. Thus, an ad-

[1]This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 813162. The content of this paper reflects the views only of their author(s). The European Commission/ Research Executive Agency are not responsible for any use that may be made of the information it contains.

versary can leverage additional records that she possesses to find a target user in the dataset, even when personal attributes are removed [6], and glean insights about them. Such insights may include:

- Whether he has an active nightlife (on what days, what times)
- When he commutes to the workplace. Whether he is active during office hours.
- When he gets back home, the time he usually leaves the house.

Needles to say, such information is extremely sensitive, and may be utilized by a wide variety of malicious actors such as thieves, burglars, stalkers, employers, etc.

Fallacies and pitfalls In order to effectively mitigate the above-mentioned threats, the following common misconceptions must be dispelled.
Fallacy: Removing participants' identifiers (name, phone number, email address, etc.) is enough to protect their privacy.
Since fitness data themselves carry a wealth of information about the person who produce them, it might be possible to identify the users based solely on their activity information.
Fallacy: Removing physical attributes (age, weight, height, etc.) fully protects privacy.
Some of the data produced by the fitness trackers depend on physical parameters of the wearer. Therefore, it may be feasible to reconstruct those characteristics.
Pitfall: Publishing all the fitness information collected by the wearable device.
The more fitness data are published, the likelier the attacker infers sensitive insights.

Guidelines for privacy protection. Besides traditional anonymization approaches, user-specific information carried by time series of fitness records (steps, calories, etc.) must be limited via the following approaches:

- *Re-sampling/sub-sampling*, which consists in reducing the amount of samples of a time series by aggregating the data at a lower frequency. For instance, the total calories burned in a day could be published instead of 24 hourly records.
- *Reducing the granularity* in a similar fashion to the generalization principle applied for microdata. This means, e.g., binning daily steps in a range $8000 - 8500$ instead of storing a more precise value such as 8361.

Finally, fitness datasets should be published with a specific purpose in mind (e.g., showing how a physical activity intervention affected certain parameters), and all the information which is irrelevant to that purpose should be discarded.

References

[1] Thambawita V, Hicks S, Borgli H, Pettersen SA, Johansen D, Johansen H, et al.. PMData: A sports logging dataset. OSF Preprints; 2020. Available from: https://osf.io/k2apb.
[2] Open Humans Fitbit Connection. OpenHumans; 2016. Available from: https://www.openhumans.org/activity/fitbit-connection/.
[3] Sweeney L. Achieving k-anonymity privacy protection using generalization and suppression. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems. 2002;10(05):571-88.
[4] Xiao X, Tao Y. Personalized privacy preservation. In: Proceedings of the 2006 ACM SIGMOD international conference on Management of data; 2006. p. 229-40.
[5] Fitbit help;. Available from: https://help.fitbit.com/articles/en_US/Help_article/1141.
[6] Marchioro T, Kazlouski A, Markatos E. User Identification from Time Series of Fitness Data. In: Proceedings of the 18th International Conference on Security and Cryptography - SECRYPT,. INSTICC. SciTePress; 2021. p. 806-11.

Challenges of Trustable AI and Added-Value on Health
B. Séroussi et al. (Eds.)

doi:10.3233/SHTI220636

Telemedicine for Diabetes in Norway

Dillys LARBI[a,1], Henriette LAUVHAUG NYBAKKE[a], Karianne F. LIND[a], Eirin RØDSETH[a] and Elia GABARRON[a,b]
[a]*Norwegian Centre for E-health Research, Tromsø, Norway*
[b]*Department of Education, ICT and Learning, Østfold University College, Halden, Norway*

Abstract. Telemedicine can be useful for diabetes patients living remotely, especially during pandemic times. We aimed to identify current knowledge of the use of telemedicine for diabetes in Norway by conducting a review of the literature. Telemedicine is mostly beneficial, and it seems that it can be adopted into the usual diabetes care in Norway as a low-cost alternative.

Keywords. Diabetes; Telemedicine; Remote consultation

1. Introduction

Telemedicine solutions can ensure continued care for diabetes patients living remotely, especially during infectious disease outbreaks [1,2]. We conducted a literature review to summarize the current evidence on the use of telemedicine for diabetes in Norway.

2. Methods

We searched in 3 scientific databases and 1 repository for relevant publications related to telemedicine and diabetes in Norway. We then extracted and summarized information about the use of telemedicine for diabetes care and the technologies used.

3. Results

Telemedicine was used to diagnose and counteract other diabetes complications [3-10], and monitor glycaemic levels [2,11-15]. Telemedicine technologies used included interactive wound platform [3-5,7,8,10], mobile phone with a self-management system [11-15], and image sharing technology [9,16].

4. Discussion

Since the use of telemedicine for the management of diabetes and its associated complications reduces HbA1c levels, minimizes the occurrence of hypoglycaemic

[1] Corresponding author, Dillys Larbi, Norwegian Centre for E-health Research, Sykehusveien 23, 9019 Tromsø, Norway; E-mail: dillys.larbi@ehealthresearch.no.

events, and improves the overall quality of life of diabetes patients [2,17], its use seems appropriate to ensure continuity of care for diabetes patients living remotely.

5. Conclusions

Telemedicine is mostly beneficial, and it could be adopted into the usual diabetes care in Norway as a low-cost alternative, especially for highly engaged individuals.

References

[1] Aberer F, et al. Application of Telemedicine in Diabetes Care: The Time is Now. Diabetes Ther. 2021;12(3):629-39.
[2] Birkeland KI. Some Lessons Learned About Diabetes and COVID-19 During the Early Stage of the Epidemic in Norway. J Diabetes Sci Technol. 2020;14(4):718-9.
[3] Iversen MM, et al. Effect of a telemedicine intervention for diabetes-related foot ulcers on health, well-being and quality of life: secondary outcomes from a cluster randomized controlled trial (DiaFOTo). BMC Endocr Disord. 2020;20(1):157.
[4] Smith-Strøm H, et al. The Effect of Telemedicine Follow-up Care on Diabetes-Related Foot Ulcers: A Cluster-Randomized Controlled Noninferiority Trial. Diabetes Care. 2018;41(1):96-103.
[5] Smith-Strøm H, et al. An integrated wound-care pathway, supported by telemedicine, and competent wound management-Essential in follow-up care of adults with diabetic foot ulcers. Int J Med Inform. 2016;94:59-66.
[6] Kolltveit BC, et al. Telemedicine in diabetes foot care delivery: health care professionals' experience. BMC Health Serv Res. 2016;16:134.
[7] Kolltveit BH,et al. Conditions for success in introducing telemedicine in diabetes foot care: a qualitative inquiry. BMC Nurs. 2017;16:2.
[8] Kolltveit BH,et al. Telemedicine follow-up facilitates more comprehensive diabetes foot ulcer care: A qualitative study in home-based and specialist health care. J Clin Nurs. 2018;27(5-6):e1134-e45.
[9] Rotvold GH, et al. Telemedicine screening for diabetic retinopathy: staff and patient satisfaction. J Telemed Telecare. 2003;9(2):109-13.
[10] Iversen MM, et al. Telemedicine Versus Standard Follow-Up Care for Diabetes-Related Foot Ulcers: Protocol for a Cluster Randomized Controlled Noninferiority Trial (DiaFOTo). JMIR Res Protoc. 2016;5(3):e148.
[11] Holmen H, et al. A Mobile Health Intervention for Self-Management and Lifestyle Change for Persons With Type 2 Diabetes, Part 2: One-Year Results From the Norwegian Randomized Controlled Trial RENEWING HEALTH. JMIR Mhealth Uhealth. 2014;2(4):e57.
[12] Torbjørnsen A, et al. A Low-Intensity Mobile Health Intervention With and Without Health Counseling for Persons With Type 2 Diabetes, Part 1: Baseline and Short-Term Results From a Randomized Controlled Trial in the Norwegian Part of RENEWING HEALTH. JMIR Mhealth Uhealth. 2014;2(4):e52.
[13] Torbjørnsen A, et al. Acceptability of an mHealth App Intervention for Persons With Type 2 Diabetes and its Associations With Initial Self-Management: Randomized Controlled Trial. JMIR Mhealth Uhealth. 2018;6(5):e125.
[14] Ribu L, et al. Low-intensity self-management intervention for persons with type 2 diabetes using a mobile phone-based diabetes diary, with and without health counseling and motivational interviewing: protocol for a randomized controlled trial. JMIR Res Protoc. 2013;2(2):e34.
[15] Arsand E, et al. Mobile phone-based self-management tools for type 2 diabetes: the few touch application. J Diabetes Sci Technol. 2010;4(2):328-36.
[16] Hernández C, et al. Integrated care services: lessons learned from the deployment of the NEXES project. Int J Integr Care. 2015;15:e006.
[17] Doupis J, et al. Smartphone-Based Technology in Diabetes Management. Diabetes Ther. 2020;11(3):607-19.

Challenges of Trustable AI and Added-Value on Health
B. Séroussi et al. (Eds.)

doi:10.3233/SHTI220637

Teaching Teledermatology: A Simulation Pilot for Health Professional Students

Blake LESSELROTH[a,1], Helen MONKMAN[b], Ryan PALMER[c], Andrew LIEW[a], Elizabeth SOO[a], Shannon IJAMS[a], Liz KOLLAJA[a], Kristin RODRIGUEZ[a], Juell HOMCO[a], Christina KENDRICK[a,d], Anna WICKHAM[c], and Frances WEN[a]
[a]*University of Oklahoma-Tulsa, School of Community Medicine, USA*
[b]*School of Health Information Sciences, University of Victoria, Canada*
[c]*Kennedy & Company, Washington DC, USA*
[d]*Tulsa Dermatology Clinic, Inc., USA*

Abstract. We developed a teledermatology simulation to give medical and physician assistant students practice with live videoconferencing and store-and-forward workflows. The simulation included (1) pre-session reading; (2) a brief teledermatology didactic; (3) a simulated encounter with a standardized patient; and (4) faculty-led debriefs. The faculty observed students during the simulation and distributed a post-session learner satisfaction survey. Although students had mixed feelings about the simulation, 88% said the workshop met or exceeded expectations.

Keywords. Telehealth, telemedicine, simulation, medical education, dermatology

1. Introduction

At the University of Oklahoma-Tulsa School of Community Medicine (OUSCM), we developed a medical curriculum on primary care telemedicine. Considering that 10% of primary care visits are for dermatologic complaints, we dedicated one workshop to teledermatology so students could practice the virtual skin exam [1]. There are few reports of simulations for teledermatology. Therefore, we developed and piloted our own scenario of a patient presenting with tinea versicolor. In this abstract, we (1) describe our simulation; (2) report preliminary findings; and (3) share lessons learned.

2. Methods

We based simulation learning objectives on telemedicine competencies published by the Association of American Medical Colleges [2]. Participants (N = 47) included third-year medical students and second-year physician assistant students.

The educational activity included didactics (i.e., dermatology in primary care and teledermatology) and a simulation adapted from work by Palmer et al. [3]. Our simulation was of a patient presenting with tinea versicolor. We gave students a short, written history and placed digital images of the simulated patient's skin condition on the workstation desktops. We expected students to review the history and images, interview standardized patients (SP) using video conference, and propose a treatment plan. We

[1] Corresponding Author, Blake LESSELROTH; E-mail: blake-lesselroth@ouhsc.edu.

conducted this activity at our simulation center, which has 10 exam rooms, each with broadband Wi-Fi; telemedicine workstations; and video conferencing software. We used a post-session survey to measure student satisfaction with the simulation [4]. We scored items on a 5-point Likert scale with anchors of "strongly disagree" to "strongly agree".

3. Results

Forty-seven students completed the module; 26 completed the post-session survey (**Figure 1**). Of those who responded, students perceived some aspects of the simulations more positively (e.g., clear objectives) than others (e.g., managing technical challenges).

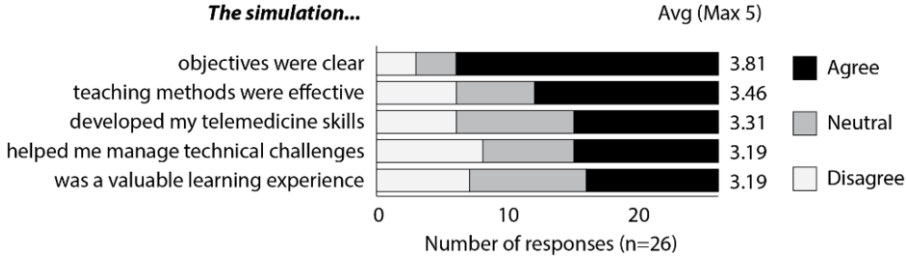


Figure 1. Results of our post-session survey.

4. Discussion

While most students completed the virtual encounter successfully, several had difficulty finding the digital images saved on the workstation. We observed students use several strategies to manage diagnostic uncertainty. For example, many proposed a working diagnosis and preliminary plan but arranged a future in-person visit to conduct additional tests. Most students found the session to be a valuable learning experience. During debriefings, many said they favored a live encounter over telemedicine when seeing patients, but most said they preferred telemedicine if seeking care as a patient. This simulation demonstrates a practical and effective strategy for teaching teledermatology.

References

[1] Farr MA, Duvic M, Joshi TP. Teledermatology During COVID-19: An Updated Review. Am J Clin Dermatol. 2021 Jul;22(4):467-475. doi: 10.1007/s40257-021-00601-y. Epub 2021 Apr 9. PMID: 33835345; PMCID: PMC8032846.

[2] AAMC, Telehealth Competencies Across the Learning Continuum, in, AAMC, Washington, DC, 2021.

[3] Palmer RT, Rdesinski RE, Galper A, Cantone R, Shaver J, Symkowick M, Biagioli FE. Assessing the impact of a telemedicine simulation on clinical learners. J Fam Med Community Health. 2017;4(5):1120..

[4] Levett-Jones T, McCoy M, Lapkin S, Noble D, Hoffman K, Dempsey J, Arthur C, Roche J. The development and psychometric testing of the Satisfaction with Simulation Experience Scale. Nurse Educ Today. 2011 Oct;31(7):705-10. doi: 10.1016/j.nedt.2011.01.004. Epub 2011 Feb 2. PMID: 21288606.

Challenges of Trustable AI and Added-Value on Health
B. Séroussi et al. (Eds.)

doi:10.3233/SHTI220638

IoT in Elementary School for Everyone - A Research Plan

Karin AHLIN[a,1], Erik WÄSTLUND[a], Awais AHMAD[b], Sture NÖJD[a], Peter MOZELIUS[c] and Ali Hassan SODHRO[d]
[a]*Service Centre Research, Karlstad University, Sweden*
[b]*Uppsala University, Sweden*
[c]*Mid Sweden University, Sweden*
[d]*Kristianstad University, Sweden*

Abstract. We propose a tentative research plan to increase students' mental health in elementary schools by implementing Internet of Things (IoT) technology. The research plan should answer how to support students' mental health using IoT solutions and the critical factors influencing testbeds for IoT solutions with the previously mentioned purpose. Our intended research method is Design Science, which we plan to use stepwise.

Keywords. Internet of Things, elementary school, research plan, design science research

1. Introduction

It can be seen today that school contributes to the decrease in children's and young people's mental health. Today, the health status of children and young people is occasionally measured through manual tools such as interviews and questionnaires. With this project, we want to find more accurate ways to measure elementary school students' health. One way of measuring is by using Internet of Things (IoT) Technology. IoT is any device connected to the Internet, usually a physical object including sensors, such as wearable devices [1]. The project aims to design, develop and evaluate IoT solutions for improving students' mental health in an elementary school. Therefore, this paper aims to present a tentative research plan for "IoT in elementary school for everyone".

2. The project

The pilot school for this project is an elementary school in the west of Sweden. The school is located in the central part of a smaller city, within walking distance to the city library and a swimming and multisport hall. The school is designed for 'home spaces' to avoid unnecessary classroom movements to reduce stress. One overall strategy for this school's staff is working with mental health among students, long-term school absences, and developing inclusive and accessible learning environments. Students' health data is today collected through physical meetings and conversations with students and their

[1] Corresponding author, Karin Ahlin, Karlstad University, Service Centre Research, Universitetsgatan 2, 651 88 Karlstad, Sweden; E-mail: Karin.ahlin@kau.se.

guardians, compiled in a regional database. This provides a basis for the school's work to reduce risk factors and strengthen protective factors for students. However, the data only shows the value from a given time and measurements for entire groups of students are also distributed over the year.

This project intends to gather and provide in-the-moment data to the school on an ongoing basis. The project's activities involve developing and testing innovative IoT ideas using the school as a testbed to reach the goal. Because schools are complex environments, it is not possible to duplicate artificially. Hence this project will be embedded in the school, with safeguards to protect the identities and welfare of staff, teachers, and children. There are few such testbeds, and by gaining knowledge, learning and experiences from the project, more actors, such as entrepreneurs, municipalities or schools, can participate in the results and use them for their activities to further develop systems for reducing risk factors and strengthening protective factors for students. Therefore, one activity in the project is to invite companies and entrepreneurs to propose IoT solutions which can be safely tested through this project in a natural school environment. The resulting benefits to business and entrepreneurship also benefit schools and students by allowing businesses to develop tools that are more likely to fit the needs of the students. With this, we want to contribute to the work with users' requirements and thereby benefit societal development.

The research methodology for this study is design science research (DSR) that consists of a five-step process: 1) to explicate the problem, 2) to define requirements for an artefact, 3) to design and develop an artefact, 4) to demonstrate the artefact and finally 5) to evaluate an artefact [2]. We will use purposive sampling [3], listening to teachers, school nurses, social educators, principals, and special educators while gathering requirements. There are several ways to listen to them, like individual or group interviews, focus groups, and workshops [3]. We can then develop prototypes based on the gathered requirements. These prototypes are expected to result in insights into possible services and work processes that give the school better conditions for promotion and preventive student health work, which contributes to strengthening the health and learning of the individual.

References

[1] Sodhro AH, Ahlin K, Ahmad A, Mozelius P, editors. Internet of Medical Things for Independent Living and Re-Learning. GLOBAL HEALTH 2021: The Tenth International Conference on Global Health Challenges, Barcelona, Spain, October 3-7, 2021; 2021

[2] Johannesson P, Perjons E. An introduction to design science. : Springer International Publishing AG; 2014.

[3] Bryman A. Samhällsvetenskapliga metoder, 3e upplagan: Liber; 2018.

Challenges of Trustable AI and Added-Value on Health
B. Séroussi et al. (Eds.)

doi:10.3233/SHTI220639

EEG Spectral Changes Linked to Psychiatric Medications: Computational Pipeline for Data Mining and Analysis

Anna MAXION[a], Arnim Johannes GAEBLER[b,c], Daniela ALBIEZ[d], Klaus MATHIAK[b], Jana ZWEERINGS[b] and Ekaterina KUTAFINA[d,e,1]

[a] *Junior Research Group Neuroscience, Interdisciplinary Center for Clinical Research Within the Faculty of Medicine, RWTH Aachen University, Aachen, Germany*
[b] *Department of Psychiatry, Psychotherapy and Psychosomatics, Medical Faculty, RWTH Aachen University, Aachen, Germany*
[c] *Institute of Physiology, Medical Faculty, RWTH Aachen University, Aachen, Germany*
[d] *Institute of Medical Informatics, Medical Faculty, RWTH Aachen University, Aachen, Germany*
[e] *Faculty of Applied Mathematics, AGH University of Science and Technology, Krakow, Poland*

Abstract. The presented computational pipeline is designed to analyze drug-induced changes in EEG data from the Temple University EEG Corpus. The data is cleaned from artifacts, pre-processed, the averaged absolute and relative frequency powers are calculated and compared to a control group. Thus, different research hypotheses can be tested with the intention to reuse accessible data collections.

Keywords. EEG, electroencephalography, antipsychotics, clozapine, risperidone, open science.

1. Introduction

Electroencephalography (EEG) is commonly used to analyze the influence of medication on brain physiology [1], since it allows to observe brain dynamics with great temporal precision, while being an economically affordable method. The Temple University EEG Corpus [2] consists of more than 30 000 data sets, and thus, provides the necessary basis to compute comparisons across large patient samples with diverse medication. The presented computational pipeline automatically searches the text reports of the full database for specified medications and downloads the corresponding EEG records for further processing leading to extraction of the spectral features. The EEG records for each drug are statistically compared to a control group. Here, preliminary results for the common antipsychotic drugs risperidone and clozapine are presented to investigate the feasibility of large-scale comparisons of different medications.

[1] Corresponding Author, Ekaterina Kutafina, Pauwelstr. 30, 52074 Aachen, Germany; E-mail: ekutafina@ukaachen.de.

2. Methods

The presented computational pipeline is implemented in MATLAB with the additional plug-in EEGLAB 2021.0 [3] for EEG data processing. It can be accessed from https://github.com/rwth-imi/Drug_induced_spectral_changes_TUH. In this project, we used the data from the main TUH EEG Corpus [2] to mine for EEG sets associated with specific medication information and TUH Abnormal Corpus (v2.0.0), which contains 1521 EEG files from 840 patients labelled as "normal" to serve as study control group. After mining for EEG files, the raw EEG data is pre-processed, the artefacts are removed, the data is divided into epochs, the averaged per channel powers of the frequency bands are computed and outliers are removed. For the statistical analysis patients using risperidone or clozapine were compared with the control group using a Welch's two sample t-test due to the largely different sample size to avoid issues with different variances.

3. Results

After pre-processing of the raw EEG data 62 patients remained from originally 277 patients in the risperidone group and 22 patients from originally 24 patients in the clozapine group. In the control group, 632 patients from originally 840 patients remained after pre-processing and exclusion of patients receiving neuroactive medication. Welch's t-tests showed an increase in global delta and theta powers and a decrease in alpha-by-theta ratio for both medications. In alpha and beta bands mixed effects were observed with overall tendency to increase in the middle central and decrease in the peripheral.

4. Discussion and Conclusions

Our results are in line with previous studies: Lee et al. [4] addressed the effect of a single dose treatment of risperidone compared to placebo, which resulted in a global increase of absolute powers in delta and theta bands. A similar result was reported by MacCrimmon et al. [5] for clozapine but being more pronounced in frontal areas. The same effect could be observed for both drugs in our data, although for clozapine the occipital delta increase is dominating. Future results of additional drug groups can be utilized to make informed decisions about the design of controlled clinical studies.

References

[1] Hyun J, Baik MJ, Kang UG. Effects of Psychotropic Drugs on Quantitative EEG among Patients with Schizophrenia-spectrum Disorders. Clin Psychopharmacol Neurosci. 2011 Aug 28;9(2):78–85.
[2] Obeid I, Picone J. The Temple University Hospital EEG Data Corpus. Front Neurosci. 2016.
[3] Delorme A, Makeig S. EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. Journal of Neuroscience Methods. 2004 Mar;134(1):9–21.
[4] Lee DY, Lee KU, Kwon JS, Jang IJ, Cho MJ, Shin SG, et al. Pharmacokinetic-pharmacodynamic modeling of risperidone effects on electroencephalography in healthy volunteers. Psychopharmacology. 1999 Jun 21;144(3):272–8.
[5] MacCrimmon D, Brunet D, Criollo M, Galin H, Lawson JS. Clozapine Augments Delta, Theta, and Right Frontal EEG Alpha Power in Schizophrenic Patients. ISRN Psychiatry. 2012 May 10;2012:e596486.

Challenges of Trustable AI and Added-Value on Health
B. Séroussi et al. (Eds.)

doi:10.3233/SHTI220640

Ambulatory Peritoneal Dialysis Analysis Framework

Nicolas INACIO[a], Gaëlle LISSORGUES[a], Adrien UGON[b,1], Corinne BAGNIS[c], Christian LOVIS[d] and Mina BJELOGRLIC[d]
[a] *UMR 9007 CNRS, ESIEE-Paris, CNAM, Gustave Eiffel University, ESYCOM, Marne-la-Vallée, France*
[b] *Sorbonne Université, CNRS, LIP6, 4, Place Jusssieu, 75005, Paris, France*
[c] *Nephrology Department, AP-HP, Hôpital Pitié-Salpêtrière, Paris, France*
[d] *Division of Medical Information Sciences, University Hospitals of Geneva and University of Geneva, Geneva, Switzerland*

Abstract. This paper presents the design of an autonomous tracking device to enhance understanding of ambulatory peritoneal dialysis. The resulting tool aims to serve as a framework for research analysis and a decision support for treatment adjustments in peritoneal dialysis.

Keywords. Peritoneal dialysis, chronic kidney disease, monitoring device, database pipeline, framework.

1. Introduction

Peritoneal dialysis is a treatment for patients with chronic kidney disease that uses the peritoneum's biological mechanism as a natural filter to replace defective kidney functions [1]. It has to be done in the hospital several times a week, which implies many constraints in the patient's life. However, this operation is facilitated if performed at home on ambulatory [2]. Despite this added comfort, ambulatory peritoneal dialysis is very difficult for caregivers to monitor [3]. The quantity of readings is not sufficient, as there is 36% risk of peritoneal degradation and loss of ultrafiltration after four years on peritoneal dialysis [4,5]. Currently, most patients requiring dialysis prefer to sacrifice their comfort for better supervised operations in hospital - only 6.2% of dialysis patients are on peritoneal dialysis while the rest are on hemodialysis [6]. This paper proposes a research analysis framework in order to simplify the follow-up of the data remotely.

2. Methods and Results

Several variables were selected to constitute the patient's data (Fig.1). For each, a preprocessing function was developed to clean the signal. In most cases, a median filter was applied - window size 5, as well as a notch filter at 50 Hz or a threshold. The EDM (Empirical Mode Decomposition) method was used to remove oscillations due to

[1] Corresponding Author, Adrien UGON; E-mail: adrien.ugon@esiee.fr.

breathing or heartbeat in the bioimpedance. By removing the uninteresting IMFs (Intrinsic Mode Function) [7], it is possible to recover a smooth signal.

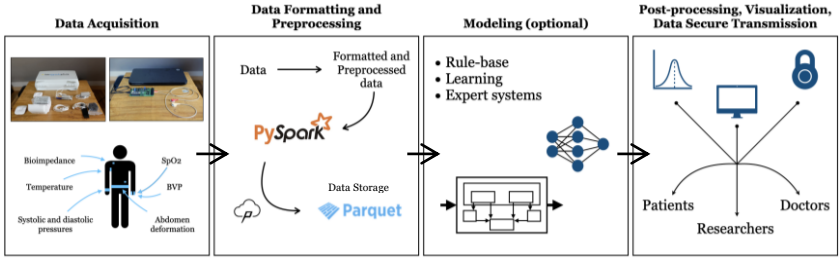


Figure 1. Data pipeline of the Ambulatory Peritoneal Dialysis Analysis Framework (SpO2 : blood oxygenation, BVP : blood volume pressure).

The database was developed in Python using the PySpark and Petastorm libraries [8,9] for structure and autonomous filling, and saved in Parquet [10]. This format is designed to accommodate a large number of entries [8,10] and to work with machine learning processes. The framework is automated from the collection of the data to its entry into the database (Fig. 1).

3.Discussion and Conclusions

This paper presents a data routing pipeline and framework to measure, store, and analyze automatically physiological parameters for continuous monitoring of ambulatory peritoneal dialysis. The feasibility of this device has been tested, but further research and additional measurement campaigns need to be organized on healthy and sick patients for comparison and complete demonstration. The work presented here aims at facilitating research from a medical point of view in order to enable ambulatory peritoneal dialysis to be better managed, offering a decision support tool for treatment adjustments.

References

[1] Mehrotra R, Devuyst O, Davies SJ, Johnson DW. The Current State of Peritoneal Dialysis. J Am Soc Nephrol. 2016 Nov;27(11):3238-3252. doi: 10.1681/ASN.2016010112. Epub 2016 Jun 23.
[2] Yang F, Luo N, Lau T, Yu ZL, Foo MWY, Griva K. Health-Related Quality of Life in Patients Treated with Continuous Ambulatory Peritoneal Dialysis and Automated Peritoneal Dialysis in Singapore. Pharmacoecon Open. 2018 Jun;2(2):203-208.
[3] Quinn RR, Oliver MJ. Is assisted peritoneal dialysis an alternative to in-center hemodialysis? Perit Dial Int. 2006 Nov-Dec;26(6):650-3. PMID: 17047230.
[4] Bellavia S, Coche E, Goffin E. Exploration d'une perte d'ultrafiltration en dialyse péritonéale [Exploration of ultrafiltration failure in peritoneal dialysis]. Nephrol Ther. 2008 Dec;4(7):590-6. French.
[5] Goffin E. Peritoneal membrane structural and functional changes during peritoneal dialysis. Semin Dial. 2008 May-Jun;21(3):258-65.
[6] Insuffisance rénale, Inserm - La science pour la santé. (n.d.). https://www.inserm.fr/information-en-sante/dossiers-information/insuffisance-renale (accessed July 16, 2021).
[7] EMD: Empirical Mode Decomposition — emd 0.4.0 documentation
[8] PySpark Documentation — PySpark 3.1.2 documentation
[9] Petastorm — petastorm 0.11.0 documentation, (n.d.). https://petastorm.readthedocs.io/en/latest/index.html (accessed July 20, 2021).
[10] Apache Parquet, (n.d.). https://parquet.apache.org/ (accessed July 20, 2021).

Challenges of Trustable AI and Added-Value on Health
B. Séroussi et al. (Eds.)

doi:10.3233/SHTI220641

Validation of Multiple Path Translation for SNOMED CT Localisation

Stefan SCHULZ[a,1], Martin BOEKER[b] and Andrea PRUNOTTO[b]
[a] *IMI, Medical University of Graz, Austria*
[b] *Institute for AI in Healthcare, Technical University of Munich, Germany*
[c] *IMBI, University of Freiburg, Germany*

Abstract. The MTP (multiple translation paths) approach supports human translators in clinical terminology localization. It exploits the results of web-based machine translation tools and generates, for a chosen target language, a scored output of translation candidates for each input terminology code. We present first results of a validation, using four SNOMED CT benchmarks and three translation engines. For German as target language, there was a significant advantage of MTP as a generator of plausible translation candidate lists, and a moderate advantage of the top-ranked MTP translation candidate over single best performing direct-translation approaches.

Keywords. Machine Translation, SNOMED CT, NLP

1. Introduction

SNOMED CT's [1] acceptance depends on localization, i.e. its adaptation to the end users' language. SNOMED CT is distributed in English and Spanish, with official translations available for some other languages. Translations provide a unique label (*fully specified name*, FSN) for each localized SNOMED CT code. As terminology translation is expensive and time-consuming, free machine translation tools bear the promise to accelerate this process. The MTP (multiple translation path) approach [2] was developed to generate, for a chosen target language, a scored output of translation candidates (TCs) for each SNOMED CT code and the related source FSNs. Besides direct translations, e.g. from Spanish to German, additional paths are created via support languages (e.g. Spanish via English to German) in order to collect a majority-vote TC list per code.

2. Material and Methods

We compared direct with MTP translations to German using four benchmarks: (i) value sets of the German BfArM Catalogue linked to SNOMED CT, enriched with synonyms provided by [3]; (ii) the unofficial translation of an early SNOMED CT version, (iii) a random subset of the 2021 SNOMED CT release, translated by medical students; and (iv) the SNOMED CT Starter Set collection [4], enriched by German synonyms. English, Spanish and Swedish were taken as source languages for Google Translator, DeepL and Systran. Danish, Dutch, Norwegian, Italian, Portuguese, Polish and Russian were chosen

[1] Corresponding Author, Stefan SCHULZ; E-mail: stefan.schulz@medunigraz.at

as support languages. From each benchmark, a random sample of 500 SNOMED CT codes that existed in the 2021 SNOMED CT release was extracted. For each code and language, the FSN (without the hierarchy tag) was fed into the process.

3. Results and Discussion

Table 1. Exact match and 1-gram BLEU metric related to benchmarks for direct and MTP translations at various rank-range choices. Uncertainties on these values were obtained by performing 10 different random choices of 500 terms (among a total of 1000 MTP-translated terms) for each benchmark, and evaluating the standard deviation both on PEM and BLEU metric.

Exact match	BfArM Catalogue	SNOMED-CT (2003)	SNOMED-CT (2021)	SNOMED-CT (Starter-set)
MTP (rank 1-5)	73.90 +/- 1.49	29.96 +/- 0.95	77.91 +/- 1.26	55.35 +/- 1.27
MTP (rank 1-4)	72.40 +/- 1.41	28.71 +/- 0.91	75.49 +/- 1.35	51.98 +/- 1.46
MTP (rank 1-3)	69.57 +/- 1.43	27.15 +/- 0.98	72.27 +/- 1.44	49.54 +/- 1.61
MTP (rank 1-2)	65.26 +/- 1.44	23.88 +/- 0.80	65.98 +/- 1.59	44.71 +/- 1.53
MTP (rank 1)	54.23 +/- 1.56	17.60 +/- 0.88	51.99 +/- 1.67	35.29 +/- 1.63
en--google--de	50.11 +/- 1.71	17.07 +/- 1.33	45.84 +/- 1.60	32.91 +/- 1.53
en--deepl--de	43.21 +/- 1.01	19.41 +/- 1.12	41.18 +/- 1.68	33.35 +/- 1.39
es--google--de	37.92 +/- 1.63	13.21 +/- 1.17	40.56 +/- 1.44	28.95 +/- 1.37
es--deepl--de	37.85 +/- 1.31	14.17 +/- 0.97	35.64 +/- 2.09	29.63 +/- 1.61
en--systran--de	37.32 +/- 1.54	14.57 +/- 1.30	29.51 +/- 1.56	29.7 +/- 1.57
sv--google--de	33.48 +/- 1.40	10.39 +/- 1.02	28.66 +/- 1.49	25.77 +/- 1.38
sv--systran--de	23.72 +/- 1.42	7.38 +/- 0.87	16.07 +/- 1.41	20.68 +/- 1.17
es--systran--de	22.31 +/- 1.23	7.72 +/- 0.94	19.10 +/- 1.31	16.03 +/- 0.82

BLEU metric	BfArM Catalogue	SNOMED-CT (2003)	SNOMED-CT (2021)	SNOMED-CT (Starter-set)
MTP (rank 1-5)	0.93 +/- 0.01	0.50 +/- 0.01	0.80 +/- 0.01	0.62 +/- 0.01
MTP (rank 1-4)	0.91 +/- 0.01	0.48 +/- 0.01	0.78 +/- 0.01	0.60 +/- 0.01
MTP (rank 1-3)	0.89 +/- 0.01	0.44 +/- 0.01	0.75 +/- 0.01	0.56 +/- 0.01
MTP (rank 1-2)	0.86 +/- 0.01	0.38 +/- 0.01	0.68 +/- 0.01	0.49 +/- 0.01
MTP (rank 1)	0.75 +/- 0.01	0.28 +/- 0.01	0.53 +/- 0.01	0.35 +/- 0.01
en--google--de	0.69 +/- 0.01	0.30 +/- 0.01	0.51 +/- 0.01	0.34 +/- 0.01
en--deepl--de	0.69 +/- 0.01	0.32 +/- 0.01	0.47 +/- 0.01	0.34 +/- 0.01
sv--google--de	0.51 +/- 0.01	0.31 +/- 0.01	0.55 +/- 0.01	0.41 +/- 0.02
en--systran--de	0.58 +/- 0.01	0.26 +/- 0.01	0.43 +/- 0.01	0.33 +/- 0.01
es--google--de	0.55 +/- 0.01	0.23 +/- 0.01	0.48 +/- 0.01	0.31 +/- 0.01
es--deepl--de	0.53 +/- 0.01	0.26 +/- 0.01	0.45 +/- 0.01	0.31 +/- 0.01
sv--systran--de	0.41 +/- 0.01	0.25 +/- 0.01	0.41 +/- 0.01	0.35 +/- 0.01
es--systran--de	0.39 +/- 0.01	0.18 +/- 0.01	0.34 +/- 0.01	0.22 +/- 0.01

For each input code, MTP resulted in 91 translation paths with various degrees of coincidence, with an average 28.7 distinct translations per code. Tab. 1 shows the comparison between MTP and direct translation (DT) in their capability to target the benchmarks, measured by % exact match (PEM) and the 1-gram BLEU metric [5]. Values strongly vary, which is explainable by different support of synonyms. MTP at rank 1 slightly outperforms any MT tool (around 1-3%). Extending to rank 1 and 2, MTP performance rises up to 20%. Including candidates at lower rank improves only to 1-3%. Restricting the comparison to rank 1 only (i.e. considering MTP a translator that produces only one translation like in DT), only the first scenario showed an advantage of MTP, whereas MTP was outperformed in the other cases particularly by the Swedish-to-German Google translation scenario by up to .06 (BLEU metric) in the fourth scenario. Finally, it is to be noted that a fraction of exact matches with the human translation is found only by MTP but not by any DT. This means that some human-like translations are found only by means of the use of intermediate languages. Our work is encouraging insofar that it suggests that for terminology translation the combination of web-based translation engines produces higher translation quality and coverage The effect size is remarkable for MTP as a shortlist creator, but still moderate when considering only top-ranked TCs.

Acknowledgements: BMBF MIRACUM, FKZ: 01ZZ1801A.

References

[1] Bodenreider O, Cornet R, Vreeman DJ. Recent Developments in Clinical Terminologies - SNOMED CT, LOINC, and RxNorm. Yearb Med Inform. 2018 Aug;27(1):129-139.

[2] Prunotto A, Schulz S, Boeker M. Automatic Generation of German Translation Candidates for SNOMED CT Textual Descriptions. Stud Health Technol Inform. 2021 May 27;281:178-182.

[3] Hashemian Nik D, Kasáč Z, Goda Z, Semlitsch A, Schulz S. Building an Experimental German User Interface Terminology Linked to SNOMED CT. Stud Health Technol Inform. 2019 Aug 21;264:153-157.

[4] SNOMED CT Starter-set. Available at: https://www.snomed.org/news-and-events/articles/snomed-ct-starter-set-translation-rfp. Accessed March 30, 2022.

[5] Papineni K, Roukos S, Ward T, Zhu WJ. Bleu: a method for automatic evaluation of machine translation. Proceedings of the 40th annual meeting of the Association for Computational Linguistics, 2002; 311-318.

Challenges of Trustable AI and Added-Value on Health
B. Séroussi et al. (Eds.)

Subject Index

This page intentionally left blank

Challenges of Trustable AI and Added-Value on Health
B. Séroussi et al. (Eds.)

Author Index