ONTOLOGY AND TEXT MINING

METHODS AND APPLICATIONS FOR HYPERTROPHIC CARDIOMYOPATHY AND BEYOND

by

LUKE THOMAS SLATER

A thesis submitted to
The University of Birmingham
for the degree of
DOCTOR OF PHILOSOPHY

Centre for Computational Biology
College of Medical and Dental Sciences
The University of Birmingham
March 2020

Dedicated to Milo Slater

"Practicing an art, no matter how well or badly, is a way to make your soul grow, for heaven's sake. Sing in the shower. Dance to the radio. Tell stories. Write a poem to a friend, even a lousy poem. Do it as well as you possibly can. You will get an enormous reward. You will have created something." – Kurt Vonnegut

ACKNOWLEDGEMENTS

I would like to thank my funders, Innovate UK, NSF, and OpenRiskNet. I would also like to acknowledge and thank my supervisors, Professor Georgios Gkoutos and Professor Simon Ball, who made this work possible.

In addition, I thank my academic colleagues John Williams, Victor Roth Cardoso, Laura Bravo, Dominic Russ, Dr Andreas Karwath, Dr Simrat Gill, Dr Tanmay Basu, Dr Saisakul Chernbumroong, Dr Animesh Acharjee, Dr Furqan Aziz, Dr Andrew Barsky, Dr Patricia Rojas, Samantha Pendleton, Dr Jie Wang, as well as my clinical colleagues Dr Amir Aziz, and Dr Dino Motti. I would especially like to thank Professor Robert Hoehndorf, for introducing me to ontology and ontology accessories, and Dr William Bradlow, who performed lots of manual data validation.

I would like to thank Maria Slater, John Slater, Hannah Slater, Matthew Slater, Milo Slater, Pud, Gus, and the rest of my family. I would also like to declare my love for Claire Martin, who also helped with proof-reading.

I would also like to express deep gratitude to my friends Benjamin Arthaud, Eric Hoftiezer, Brian Mullan, Sebastian With Olsen, Adam Goodwin, Ellie Bartoli-Leonard, Alex Brown, Shuaib Farah, Ryan Pridgeon, Shay Bacon, Eddie Curtis, Orion Sawchuck, Ben Handel, Philip Butler, Jean-Paul Grund, William Starling, David Marsden, Ed Morgan, Frank Polyn, Tom Wilkinson, Joseph Davenport, Kevin Kane, and Vez Kirkpatrick.

Finally, thank you to the many researchers, supervisors, colleagues, family, and friends, whose work, love, and support also transitively enabled this thesis.

Abstract

In this thesis we describe a number of contributions across the deeply interlinked domains of ontology, text mining, and prognostic modelling. We explore and evaluate ontology interoperability, and develop new methods for synonym expansion and negation detection in biomedical text. In addition to evaluating these pieces of work individually, we use them to form the basis of a text mining pipeline that can identify and phenotype patients across a clinical text record, which is used to reveal hundreds of University Hospitals Birmingham patients diagnosed with hypertrophic cardiomyopathy who are unknown to the specialist clinic. The work culminates in the text mining results being used to enable prognostic modelling of complication development in patients with hypertrophic cardiomyopathy, finding that routine blood markers, in addition to already well known variables, are powerful predictors.

CONTENTS

CHAPTER 1

INTRODUCTION

Every science produces information, and major tasks arise at the intersection between the science of information and science producing information. This is particularly true of biomedical science, whose data are particularly voluminous and are produced in multiple contexts: for example, by literature, clinical practice, animal experimentation, and clinical trials. Information is also produced in many modes: for example, strucured Electronic Healthcare Record (EHR) data, imaging data, background knowledge in ontologies, and clinical narrative text. These data are analysed, and the results in turn become information, stored in databases and literature.

Importantly, these multi-modal data often concern the same entities, or entities that are semantically linked. For example, a hospital's EHR contains information, encoded into various modalities, about patients that visit it. These include the diseases they suffer, treatments they undergo, their socio-economic background, their genetics, and more. By virtue of having this information about patients at the hospital, we also have information about the entities associated with these patients; about the diseases, treatments, socio-economic backgrounds, and genetics themselves. To integrate this information is to gain the opportunity to create new knowledge that was not obtainable through analysis of one source alone. In the biomedical, and particularly the medical field, this is important because it can lead to direct impacts on patient outcomes.

EHRs are in ubiquitous use throughout all modern healthcare systems, both in primary

and secondary care[24]. More recently, EHR data has become available to researchers for both specific and general uses, as the potential value of the data they contain is increasingly realised among clinical and research communities[153]. However, much of the data generated by clinical practices are contained within unstructured resources. While EHR systems contain structured information, medicine continues to be performed largely via textual communication. Most information concerning the clinical practice is stored in narratives, physician's notes, MRI reports, paper-based prescriptions, and more[54, 127].

In this thesis, we explore several technologies along a pipeline for information extraction, integration, and analysis, with a focus on ontologies and biomedical text mining in a medical setting. We describe, implement, and evaluate an ontology-based approach to text-mining, and demonstrate that this can be used to enhance information stored in a structured EHR system, and thereby enable analysis. In particular, we will explore the following subjects. First, a literature review will discuss the major concepts and research areas involved in the thesis: The Semantic Web and Computational Ontology, Healthcare Data, Text Mining, Risk Modelling, and Hypertrophic Cardiomyopathy. We follow this with an investigation into the biomedical ontology ecosystem and its methods of class inclusion and re-use, exploring problems with Minimum Information to Reference an External Ontology Term (MIREOT), and presenting an algorithm to automatically repair them, with some analysis of root causes. We then present a method of leveraging ontology term re-use and redundancy for the expansion of text mining vocabularies, showing that the approach increases the recall of information extraction and retrieval tasks. We also present a new method for detecting uncertainty and negation in text, and show in an evaluation on multiple datasets that it out-performs current state-of-the-art methods.

These approaches and algorithms are then consolidated into a text mining pipeline for patient phenotyping across a clinical text record. This is used to discover and phenotype HCM patients managed by the hospital, but unknown to the specialist clinic. Using these data, in combination with structured healthcare information, we develop a prognostic model to predict the likelihood of HCM patients developing heart failure and atrial

fibrillation over a three year period.

CHAPTER 2

LITERATURE REVIEW

2.1 The Semantic Web and Computational Ontology

Thn semantic web is a set of technologies, methodologies, and standards developed for the purpose of enabling computable expression, transmission, and analysis of data over the World Wide Web (WWW)[28]. It is defined in contradistinction to the syntactic web, which makes up the WWW that humans interact with. Most often expressed with HyperText Markup Language (HTML), the syntactic web is formulated from documents that contain natural language and image content, alongside auxiliary components such as navigation and interactive features. Importantly, the meaning of this content is derived via offline consensus.

Since it is expressed with human language, the syntactic web inherits the problems of human language. In this case, particularly, that its semantics are ephemeral, and meaning is derived by a complex social and cultural process that cannot be explicitly defined in a computable form. For example, the first paragraph of the Wikipedia article for heart failure (HF) is as follows (with citation markers removed)[15]:

> Heart failure (HF), also known as congestive heart failure (CHF) and congestive cardiac failure (CCF), is when the heart is unable to pump sufficiently to maintain blood flow to meet the body's needs. Signs and symptoms of heart failure commonly include shortness of breath, excessive tiredness, and leg swelling. The shortness of breath is usually worse with exercise or while lying down, and may wake the person at night. A limited ability to exercise is

> also a common feature. Chest pain, including angina, does not typically occur due to heart failure.

This is an example of a piece of information as expressed by the syntactic web. As humans read it, we are able to derive a large number of facts about the subject through perception, inference, and deduction. We do this through an almost automatic cognitive process, correlating and integrating the information given with information that we already know. Interestingly, the paragraph does not explicitly state that heart failure is a disease. Had a reader not already heard of the condition, they would be able to infer this information from the paragraph. This is because we would have heard of things like "excessive tiredness" and "shortness of breath," and would match the description of HF with our general idea of a disease as being when something is wrong in the body. Indeed, one of the definitions for disease given by the Cambridge Dictionary defines it as "a condition of a person, animal, or plant in which its body or structure is harmed because an organ or part is unable to work as it usually does; an illness[2]."

A computer cannot easily do this. It does not know what the concepts described by the words mean, and it does not have background knowledge to relate them to. It would not, like a human, be able to look up the meaning of 'disease' in a dictionary, or look up any of the citations given in the paragraph, because it would not understand the natural language used in them. A text mining application could be created to extract information from the article, but it would not (without semantic technologies) itself have any sense of what a disease is: the creator would have to express this knowledge in the programming logic, and interpret the meaning from its output.

The semantic web, on the other hand, necessarily expresses information in a computable format, using the linked data paradigm. Data is provided in a structured formulation, annotated with identifiers that express its semantics. By doing this, it makes explicit what is implicit in natural language data, reducing ambiguity and providing a reference point for computable understanding. DBPedia is a semantic web project, which contains linked data mined from Wikipedia sidebars for structured information, convert-

ing it into semantic data[21]. Figure 2.1 lists a selection of information concerning heart failure in DBPedia[4].

```
dbo:icd10
   I50

dct:subject
   dbc:Organ_failure
   dbc:Aging-associated_diseases
   dbc:Heart_diseases

rdf:type
   dbo:Disease
   wikidata:Q12136
```

Figure 2.1: A selection of triples from the Heart Failure DBPedia entry.

Each entry describes a relationship between the subject, heart failure, and an object. For example, one relationship is *rdf:type*, wherein the subject is heart failure, and the object is *dbo:disease*. *dbo:disease* is an identifier that links to another database entry, and in turn is linked to every other database entry that contains a link to *dbo:disease*. Importantly, we also specify what the relationship is; in this case *rdf:type*, meaning that the subject is a type of the given object.

In the previous example we have shown that by looking at the DBPedia data, we can link HF with other concepts that share the same relationship with 'disease.' But how can the computer gain an understanding of what a disease is? To gain an implicit understanding of what a disease is, we can further explore the database. By examining concepts that are of the type *dbo:disease*, we would identify the things these linked concepts, in turn, stand in relation with. For example, via the *dct:subject* relationship, we find transitive connections to subjects such as organ failure and ageing.

Explicitly, we can examine relationships to sources that contain more structured information about diseases. The other *rdf:type* relationship associated with HF is a link to another database, Wikidata[169]. The Wikidata entry contains an additional wealth of information, expressed as relationships to other objects[3]. For example, we know from

the *has effect* relationship, that diseases, albeit rarely, can cause *dying*, and that it is the *opposite of health*. Moreover, all of the relationships and objects pertaining to disease, are also in turn described by the knowledge as subjects in terms of their relationships with other objects, that can be explored in exactly the same way.

Another external database link shown in Figure 2.1 is to a medical terminology. The *dbo:icd10* relationship states that the concept expressed by this database entry is semantically equivalent to the ICD-10 code with the identifier *I50*. ICD-10[121] is a medical terminology which is frequently used in practical medical environments to establish relationships between patients and diseases. For example, databases describing patients suffering heart failure might have a relationship with *I50*, which would in turn be linked to the other sources of information we have concerning HF, including DBPedia and Wikidata.

This annotation of individuals bridges the gap between semantic web models defining attributes of things and defining *instances* of things. Now, when we have a patient annotated with this code *I50*, we can link to databases like Wikidata to automatically understand that the patient has HF, and to come to an understanding of what it means to have HF. In turn, the patient contributes to the total understanding of the disease in the knowledge graph, contributing a phenotypic profile of someone who suffers HF: their symptoms, their average blood pressure, imaging reports, outcomes, and more. This example shows how, using the linked data paradigm, we can create linked networks of information that define things with respect to their relationships with other things. In the rest of this section, we will discuss the technologies that are used in the semantic web.

2.1.1 Resource Description Framework

The Resource Description Framework (RDF) is a method of modelling and describing concepts[112]. An RDF model constructs a knowledge graph using a series of triples:

Subject The object being described.

Predicate The property of the object being described.

Object The value assigned to the given predicate in relation to the subject being described.

For the heart failure example in the introduction to this section, the Wikidata entry expresses a relationship with the subject *heart failure*, the predicate *opposite of* and the object *health*. Predicates and objects can also serve as subjects, and in this way the knowledge graph elaborates on what it means for heart failure to be the opposite of health, by building up relationships that define what each constituent part of the relationship means - just as this relationship contributes to understanding of *health*, *opposite of*, and *heart failure*. Figure 2.2 shows an example RDF graph, using some of the example relationships described so far.

While the examples given so far use literal string representations of objects, they are actually identified using Uniform Resource Identifiers (URIs). For example, the *health* object is identified using `https://www.wikidata.org/wiki/Q12147`. In this case, the URI resolves in a web browser to a page describing the object, but they do not necessarily have to resolve anywhere: their function is as a unique identifier for the concept, that can be referred to anywhere. They are also often shortened to identifiers, comprised of a conventional name for the database and a unique code identifying the concept in that database, e.g. *wikidata:Q12136* for *heart failure*. Standard vocabularies, such as Dublin Core[1], define predicates and objects widely used by convention throughout RDF databases. Some of the standard relationships provide human-readable metadata, such as the labels given in the previous examples.

2.1.2 Ontologies

RDF provides a framework for describing objects, while ontologies describe the kinds of objects that can exist in a domain, and define a formal logic basis for their semantics. Computational ontologies are inspired by the philosophical ontology: a consideration of

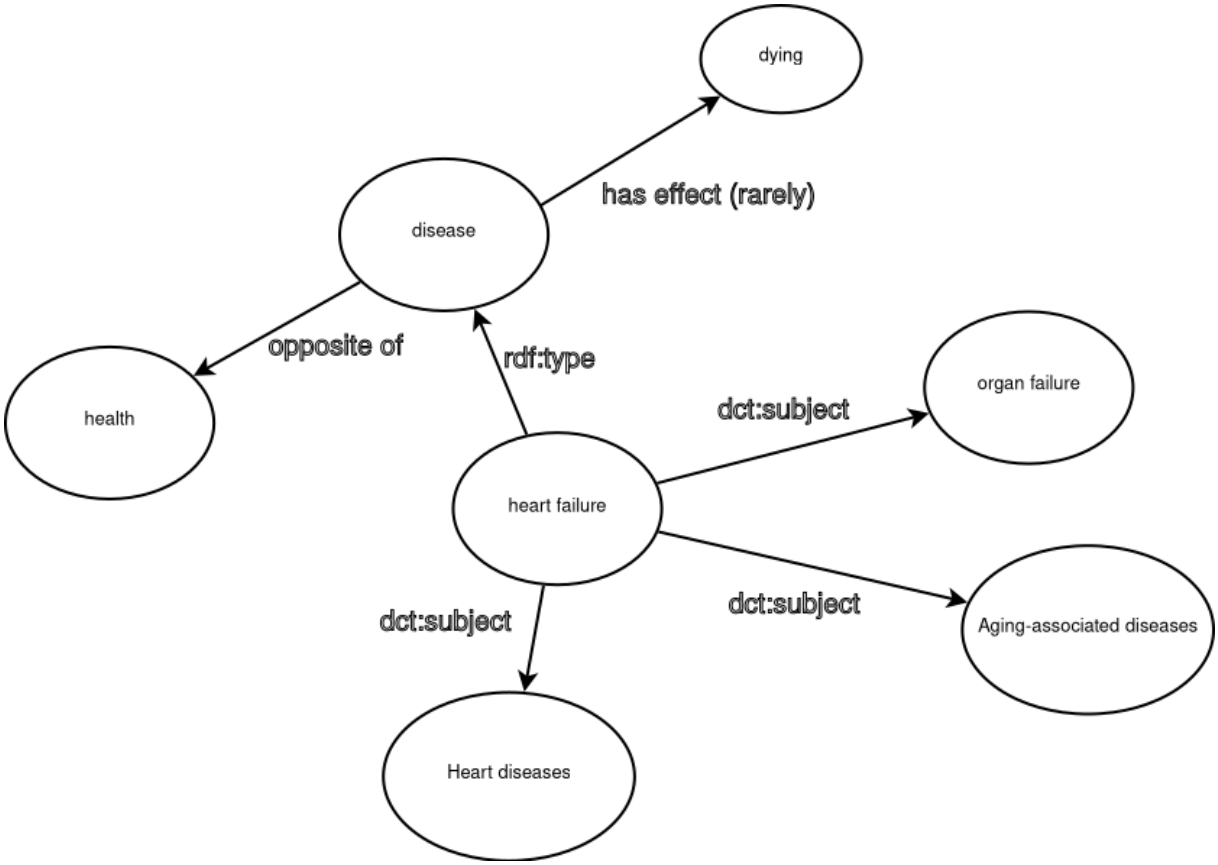

Figure 2.2: An example knowledge graph built from triples that describe relationships for heart failure and associated concepts in Wikidata and DBPedia. For ease of understanding, subjects and objects are given with their text labels. In reality, these concepts would be identified by their unique URLs, and the labels would be specified using an additional triple relationship.

things in the world, and how they relate to each other[74]. Computational ontologies have a long history of study, growing from the tradition of analytic philosopy and formal logic. Previous to RDF they mostly took the form of expert systems or Prolog knowledgebases[144]. In their modern form, they have grown from an effort to extend RDF models with formal semantics. These efforts began with RDF Schema (RDFS), which defines a list of standard properties for RDF that together form the basis for an ontological organisation of concepts modelled using the language[10]. Referring back to the heart failure example, Wikidata uses one of these properties, *rdfs:subclassOf* to form a class hierarchy, a subset of which is shown in Figure 2.3.

The RDFS specification defines the *rdfs:subClassOf* predicate as follows:

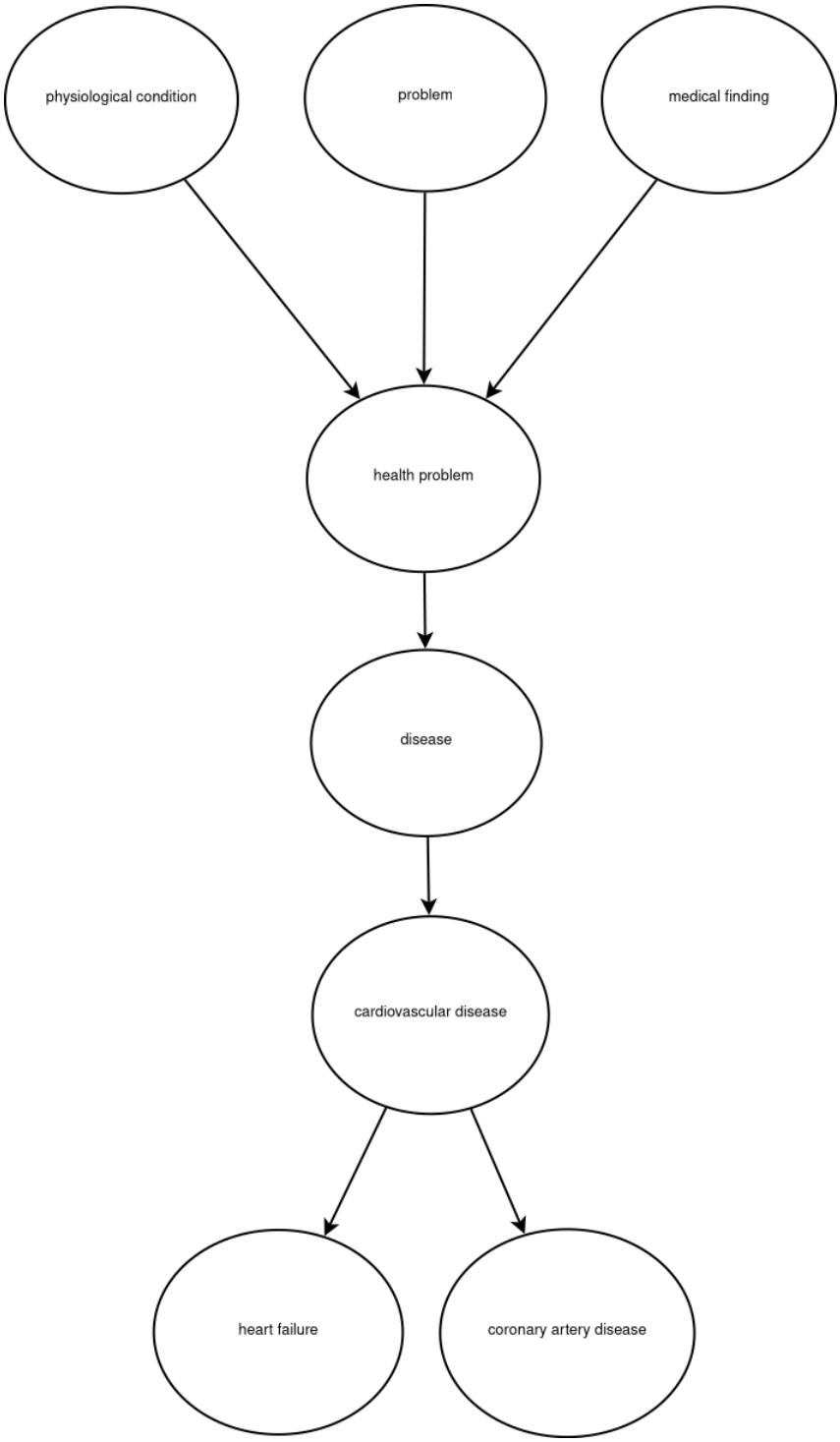

Figure 2.3: Part of the class hierarchy described by Wikidata for cardiovascular diseases. Each directional relationship is formed from a triple that describes an rdfs:subClassOf relationship.

> "The property rdfs:subClassOf is an instance of rdf:Property that is used to state that all the instances of one class are instances of another."

These relationships are transitive, meaning that if concept A is a subclass of concept B, and concept B is a subclass of concept C, A is also a subclass of concept C. This transitive property means that the subClassOf predicate can be used to formulate a hierarchy of concepts, wherein all instances of objects are also instances of all objects their parent concepts are instances of.

This logical formulation is an implementation of a syllogism: an argument that enables a deductive inference. Aristotle and Porphyry relate an example[19]:

"All men are mortal; Socrates is a man, therefore Socrates is mortal."

In the knowledge graph described by Figure 2.3, arguments of the same form can be made through explication of subClassOf relationships: "heart failure is a disease; a disease is a health problem, therefore heart failure is a health problem." By leveraging semantic relationships described using RDFS, automated reasoners can infer new knowledge from the explicit information encoded in the knowledge graph, using rules such as syllogism. Automated reasoners work by determining the logical consequence of every explicit assertion in a knowledgebase. For RDFS, several automated reasoners exist, such as Jena, RDFox, and GraphDB. Inferences for RDFS ontologies can be made using forward and backward chaining rule-based reasoners, which are algorithms for solving propositional logic formulae. GraphDB, for example, uses a forward-chaining reasoner to infer knowledge from RDFS ontologies[91].

Extending some of the concepts from RDFS, The Web Ontology Language (OWL) is a family of languages defined in terms of different subsets of description logics, which are in turn a fragment of first order logic. While OWL ontologies are often still expressed in RDF, the difference is that it uses a specification that expresses certain kinds of concepts and relationships between concepts that constitute an ontology. The majority of biomedical ontologies are now described with the Web Ontology Language (OWL)[170], while instance data is still mostly expressed using non-OWL RDF. OWL Ontologies can also be described in a number of other functionally equivalent languages, such as functional

syntax, turtle[31], and JSON. The OBO file format is an ontology language heavily inspired by OWL, with a more idiomatic synax. Most of its constructs can be mapped to OWL, and many ontology tools can work with both formats[162]. In addition, there are some ontologies that don't use the OWL specification, such as Prolog knowledgebases and other RDF specifications. In a functional discussion of biomedical ontologies and their composition, Hoehndorf et al. notes that, while an exact definition for computational ontology is elusive, most share four features[72]:

- Classes and relations
- Domain vocabulary
- Metadata and descriptions
- Axioms and formal definitions

Artefacts with these features describe an understanding of a particular domain in a form that can be leveraged by both humans and machines. The domain vocabulary, metadata and decriptions provide human-readable and definable knowledge, while the clases, relations, axioms, and formal definitions attach these syntactic constructs to formal semantics, which can be understood with respect to their logical restrictions and relationships with other concepts.

Manchester OWL Syntax is a human-readable format for description logic axioms. The Manchester OWL Syntax formulation of an axiom for *hypertension* (HP:000822) is shown in Figure 2.4. It defines hypertension in terms of its relationship to other qualities, anatomical strctures, and temporal constructs.

Some medical terminologies, such as SNOMED and ICD, are not considered to be ontologies because they do not provide axioms or formal logic definitions[37]. However, they use a taxonomic structure and many of the metadata features of ontologies, so they are often used in the same way and discussed in the same contexts.

```
'has part' some
    'increased pressure'
  and
    'inheres in' some
        blood
      and
        'part_of' some 'blood vessel'
  and
    'has modifier' some abnormal
  and
    'has modifier' some chronic
```

Figure 2.4: Axiom for *hypertension* (HP:000822) as defined by HPO, rendered in Manchester OWL Syntax.

Automated Reasoners

Because OWL ontologies are expressed by description logics, automated reasoners can be employed to evaluate all of the impliciations of the explicit assertions made in the ontology to determine that it is consistent (contains no contradictions). They can be applied to ontologies expressed in this family of languages through reduction of the problem of ontology entailment in OWL to one of knowledgebase satisfiability, most of the time using an analytic tableaux method. As well as consistency, automated reasoners also reveal axioms and relationships between classes that are inferred from the explicitly asserted axioms.

While many ontologies contain formal axiomatisations of their concepts, classification is costly in terms of computational complexity. Particularly in time complexity: the full OWL-DL language classification is NEXPTIME-complete. Different OWL language profiles have different time complexity profiles. Of particular note is the EL fragment, which is guaranteed to be classified in polynomial time. EL reasoners, such as ELK, are frequently used for this reason. More expressive ontologies can also be classified using EL reasoners, ignoring any axioms outside of the EL fragment[86]. Much of the previous work into large-scale reasoning over ontologies has focused on the corpus of ontologies available from BioPortal, since it is one of the largest collections of freely available ontologies[38].

Modularisation techniques have discovered locality-based modules existing within ontologies and demonstrate that for most ontologies and types of query relatively small modules can be found. Using these techniques can improve reasoner-based query performance[46]. One investigation found that the performance of certain reasoners over the set of BioPortal ontologies can be reliably predicted. The same work performed an extensive evaluation on the average classification time for each ontology[135].

Description Logic Queries

Reasoners also enable description logic queries. A complex class description, often described using Manchester OWL Syntax, can be posed to the reasoner as a new class, and the reasoner return any classes that have been inferred to satisfy that description, as subclasses, superclasses, or equivalent classes. For example, the Manchester OWL Syntax description of hypertension shown in Figure 2.4 contains several restrictions for what it means for an entity, in this context a phenotype, to be hypertension. Other phenotypes may share these components. For example, if we wanted to find other phenotypes that involved 'increased pressure', we could pose the following equivalency query:

'has part' some 'increased pressure'

This would, amongst the other subclasses of hypertension (which inherit its logical definition), return *Elevated pulmonary artery pressure* (HP:0004890). Interestingly, this condition is often known as 'pulmonary hypertension', which would then be a subclass of hypertension. However, in HP this is not the case, perhaps due to a particular design or expert decision. By using the description logic definitions, however, we are nevertheless able to explore what these concepts share in terms of their fundamental definition.

Unsatisfiability

In OWL ontologies, an unsatisfiable class is one for which there cannot be an instance while maintaining the consistency (non-contradiction) of the axioms in the ontology. For

example, if our ontology contains the axioms:

1. Something that is a phenotype is disjoint from something that is a disease.
2. Hypertension is a disease.
3. Hypertension is a phenotype.

Our first axiom describes a restriction on what a particular thing can be: that anything which is a disease, cannot also be a phenotype, and vice-versa. This is an extension of the principle of non-contradiction: that a proposition cannot be both true and false simultaneously. With the subsequent axioms, we describe what it means for something to be hypertension. That is, that this entity is a disease, and also that it is a phenotype. While the contradiction here is apparent, it is important to note that we have not created an inconsistency itself; as far as the ontology itself is concerned, there are models of the ontology that satisfy all of the axioms: one in which there are no instances of hypertension. We have, however, created an unsatisfiable class. We have created a definition for a kind of thing that cannot exist, and therefore its class description cannot be satisfied. The ontology would become inconsistent if we created an instance of something that is unsatisfiable. For example, if we were to annotate a patient with hypertension using this ontology, the ontology would become inconsistent, we would be creating an interpretation of the world wherein hypertension exists; this patient, with hypertension, has something which is a disease, and is a phenotype, while our disjointness axiom excludes this possibility.

An ontology is called incoherent if it contains any unsatisfiable classes. An inconsistent ontology is one which cannot have any model: this is usually because it contains an instance of an unsatisfiable class. There are some intentional uses of unsatisfiability, most often these are deprecated classes, which have been retired and superceded, and should not be used anymore, and should therefore no longer have instances. This can happen because a term was found to be synonymous with another term, or because it did not provide a good representation of the entity it attempts to describe. To do this, a disjointness axiom is introduced between all deprecated classes, and owl:Thing, meaning

that any instance cannot also be an instance of any of the deprecated classes. For example, the class *congenital neutropenia* (HP:0005549) was made obsolete, and replaced by a recommendation for annotators to use *neutropenia* (HP:0001875) with the modifier *congenital onset* (HP:0003577). Implicit unsatisfiable classes inferred by a disjointness axiom or other object restriction are also functional examples of unsatisfiability. For example, GO contains a disjointness axiom between *whole membrane* (GO:0098805) and *membrane region* (GO:0098589). From this the reasoner creates an implicitly unsatisfiable class description, and any instance that satisfies that class description would make the ontology inconsistent. For example, an annotation for a piece of experimental data that labels something as both a *chitosome membrane* (GO:0030661) and also an *annulate lamellae* (GO:0005642), would satisfy that inferred unsatisfiable class, making the ontology inconsistent. Otherwise, the existence of visible unsatisfiable classes in an ontology indicates that the ontology is incoherent, and where this is not intentional it is conceptually equivalent to the ontology being inconsistent. While we do not usually explicitly create instances of classes in the ontology itself, when we annotate, access, and integrate data partaking in unsatisfiable entities, we are acknowledging a model of the ontology wherein this entity is realised. That is, we are using it in a way that necessitates its inconsistency. When an ontology is inconsistent, we cannot infer anything useful from it, since any model of the ontology must necessarily accept that A is equal to NOT A. According to the principle of explosion, if this is the case then the truth value of any proposition is simultaneously true and false, and any fact can be inferred[51].

Ontology development tools and build processes include methods for checking the consistency, satisfiability, and coherency of ontologies. This is achieved by integration of an automated reasoner, which can check for consistency and mark any unsatisfiable classes. More recently, automated reasoners can provide explanations for why a class is unsatisfiable, exposing the trail of axioms leading to the root cause of the inference. There have been many efforts to develop tools to explain causes of unsatisfiability, based on an understanding of class unsatisfiability in the context of the hitting set problem.

For example, Reiter's Hitting Set Tree (HST) algorithm can obtain all justifications for OWL entailments, and therefore the axioms involved in class unsatisfiability[85]. Using the information from these algorithms, several pieces of software now exist to relay this information in an understandable manner, conducive to human-aided resolution of the logical inconsistency. These are often integrated directly into ontology development tools such as Protégé. These tools have two limitations. First, that they cannot actually highlight the true or 'correct' cause of the unsatisfiability, and that they cannot tractably retrieve large numbers of explanations[84]. The running time for the HST algorithm, for example, does not have a practical upper bound, and runs exponential to the size of the conflicting sets of classes considered[98]. Furthermore, as ontologies become more complex, the reasons for class unsatisfiability have become more complex, potentially spread across a large number of concepts and obscured by unclear inference steps performed by the automated reasoner.

Data Annotation

While ontologies describe the kinds of things in a domain, we can also link these descriptions to data about instances of these things. By annotating an entity, we are saying that it is an instance of the class. Although previous ontology examples have given the natural language labels of ontology classes, they are actually uniquely identified by Internationalised Resource Identifiers (IRIs). This is the same method used by RDF datasets, and therefore the components of RDF triples, can be linked to ontology classes via their IRI.

Non-RDF databases are also annotated using ontologies. This can take the form of full IRI references, but more often they use a shortened term identifier, which can be transformed to the full IRI with a series of string operations. OBO ontologies use a prefix identifier for the ontology followed by a colon and then a numeric identifier. For example, the term identifier (OBI:1110108) refers to a class with the full IRI *http://purl.obolibrary.org/obo/OBI_1110108* in the Ontology for Biomedical Investigations (OBI)[22]. Non-ontology terminologies, such as ICD or SNOMED also define term

identifiers, and these are also used to annotate instance data. The identifiers are constructed taxonomically (each character represents one level of the hierarchy), and so more specific or general annotations can more easily be chosen without knowledge of the taxonomic structure being referenced.

Ontology Reuse and MIREOT

Referencing and extending classes from other biomedical ontologies is common practice. We have previously discussed how ontologies create meaning for concepts by placing them in the context of relationships with other concepts, and the relationships and restrictions that they transitively inherit. When creating a description of a domain, more meaning and more information can be encoded by defining a concept in relation to concepts in other domains, as well as its own.

Upper level ontologies describe high-level or very general concepts, which are used by more specific ontologies to place their terms into a common context. This enables integration across domains. The Basic Formal Ontology (BFO) provides a metaphysical basis for the description of biomedical concepts, starting with a fundamental distinction between continuants and occurrents: essentially a distinction between material and temporal objects[145]. The Open Biomedical Ontologies Foundry is a collection of ontologies that use the same design principles, and they all use BFO as a base[160]. They also use the Relation Ontology (RO), which defines standard object properties that can be used to define relationships between objects[38]. Other ontologies outside of this collection also make heavy use of BFO and RO, in a less strict fashion. Using these standard definitions of high-level concepts and relationships allows for consistency between ontologies. These high-level ontologies fulfill the same purpose for ontology objects that ontology objects do for biomedical data: they describe a semantic equivalency between different instances of the same thing.

There are other ontologies that exist to define general concepts that see heavy re-use across domains. For example, the Units Ontology (UO) defines units of measurement

across science, the Phenotype And Trait Ontology (PATO) defines traits and qualities, and Uber-anatomy Ontology (UBERON) describes anatomy[58, 56, 113]. We have previously discussed the HPO class *hypertension* (HP:000822). HPO is an ontology that defines phenotypes, but its formal descriptions define its classes in terms of their relationships with different kinds of concepts. Figure 2.4 builds a logical definition for hypertension with reference to different anotomical, temporal, and qualitative concepts. For example, the *increased pressure* (PATO:0001576) comes from the PATO, while *blood vessel* (UBERON:0000178) comes from UBERON.

If concepts in different domains define things in terms of the same qualities and concepts, information can be integrated across domains. While HP defines human phenotypes, the Mammalian Phenotype Ontology (MP) is used primarily for mice, and the results of mice experiments are annotated using this ontology. While mice are similar to humans, which is why they are widely used as a model organism, there are still substantial differences in the conceptualisation of the human and mouse phenomes. While the class for *hypertension* (MP:0000231) in MP happens to have the same label as the HP concept, this is not guaranteed. Furthermore, there is no guarantee that classes with the same labels are semantically equivalent. Previously we discussed that in the HP conceptualisation of the human phenome, the class *Elevated pulmonary artery pressure* (HP:0004890) is used to describe what is normally known as pulmonary hypertension. Indeed, this is the case in MP, with the class *pulmonary hypertension* (MP:0003548). However, because the same ontologies are used to develop axiomatic descriptions of these concepts, these can be used to come to an understanding of the functions they share, and their semantic equivalence. The pulmonary hypertension and increased elevated pulmonary artery pressure classes share a fundamental part of their axiomatisation, that links them to the anatomy and qualities they concern, as shown in Figure 2.5. If an ontology was created with all of the axioms from HP and MP, this axiom could be passed as a description logic query that would be satisfied by both *Elevated pulmonary artery pressure* (HP:0004890) and *pulmonary hypertension* (MP:0003548). Even though the 'syntax' that humans use

to refer to and understand these concepts in the respective domains differs, we can use their axioms to connect and integrate them, and in turn the entities that are annotated with them.

```
'has part' some
    'increased pressure'
  and
    'inheres in' some
        blood
      and
          part_of some 'pulmonary artery'
        and
          'has modifier' some abnormal
```

Figure 2.5: Partial Manchester OWL Syntax axiom shared by *Elevated pulmonary artery pressure* (HP:0004890) and *pulmonary hypertension* (MP:0003548).

Another definition for hypertension comes from the Disease Ontology (DO). *hypertension* (HP:000822) describes a phenotypic abnormality: "The presence of chronic increased pressure in the systemic arterial system." DO, on the other hand, defines *hypertension* (DOID:10763) as a subclass of *artery disease* (DOID:0050828): "An artery disease characterized by chronic elevated blood pressure in the arteries." Hypertension is both a disease and a phenotype, and an annotation to either DO or HP depends on the context of the dataset.

Ontologies also exist along a continuum of specificity. While HP is a very general ontology, which aims to define most phenotypes in the human phenome, there are other ontologies which describe certain phenotypic sub-domains to a greater level of granularity. The Hypertension Ontology (HTN) defines concepts in the domain of hypertension in a more specific way. For example, it makes a distinction between elevated diastolic and elevasted systolic phenotypes, which are perhaps relevant for studies that concern hypertension particularly, but would potentially be too specific for phenotype studies concerning the whole phenome. However, because it builds on concepts from other ontologies, the information it adds it can be related to other ontologies. It re-uses the hypertension classes from HP and DOID as a base, and defines its more specific phenotypes and related entities

with respect to them. Therefore, annotations using HTN can be related back to DOID and HP, and vice-versa. In fact, HTN also includes an axiomatisation of the distinction between the HP and DOID hypertension classes we explained earlier. In its re-use of the DOID class, it has added an additional axiom, where hypertension refers to the HP concept:

'disease arises from feature' some Hypertension

This additional ontology, then, gives us a greater understanding of the concepts than we have from HP and DOID alone, because they tell us how the concepts in those ontologies relate to each other. Furthermore, the object property used to define the relationship, 'disease arises from feature' (RO:0004022) is defined in RO, providing more contextual knowledge to automated reasoners. These examples show that through ontology re-use, ontologies that describe different domains or the same domains in different contexts, can still support integration of knowledge and information between them.

We have discussed a few ways that ontologies can make use of concepts defined elsewhere, and given some of the reasons that this can be useful. There are, however, different methodologies for concept re-use. Ontologies that re-use BFO, such as those contained in the OBO Foundry, use the owl:imports closure, wherein one can provide a path to another ontology to import into the current ontology, in its entirety, upon loading. This method is often used when ontologies make heavy and general use of the concepts in the sourced ontology. For example, BFO is often imported directly: it is a minimal ontology with 36 concepts, which form the fundamental framework around which the ontologies that use it are constructed. However, this is not always sensible or feasible. Even when an ontology makes heavy use of another, only a subset of the sourced ontology is likely to be relevant. Aside from a matter of cleanliness, it also becomes a performance issue. Loading the ontology will take longer, especially since the ontologies are often downloaded from the Internet, upon loading, and classification will take longer - as all of the concepts, even those irrelevant to the concepts used will be evaluated. For phenotype ontologies explicating on one or few descriptions from large general domain ontologies, creating a

specific ontology for a particular sub-context, it does not make sense to import the entire ontology. HTN expands upon the hypertension concepts in HP and DOID, but is not concerned with any terms in those ontologies beside those directly related to hypertension. To import the whole ontology would lead to unnecessary confusion and usability issues. Besides problems of design and usability, it can quickly become impossible to work with combined ontologies because of limited hardware resources and highly complex ontologies being sourced. For several particularly large ontologies, it is already unfeasible to develop them using desktop tools like Protégé alone. Particularly for application ontologies, such as EFO[99], problems with its size are compounded by the large number of imports it would have to make.

Within the biomedical ontology community, minimum information guidelines named MIREOT (Minimum Information to Reference an External Ontology Term) were developed to support ontology reuse[38], specifically taking into account the needs of the biomedical community. Using these methods avoid the overheads involved in importing complete ontologies. Additionally, the explicit aim of the method is also to prevent inconsistency and unintended inferences by encouraging the reuse of classes that are already well-defined and established within the domain, and many biomedical ontologies make use of the MIREOT method to achieve these aims. According to the MIREOT system, including a class from an external ontology requires:

Source ontology IRI The IRI of the ontology which contains the class being included.

Source class IRI The IRI of the class to import, as given in the external ontology.

Direct Superclass IRI The IRI of the direct superclass for the imported class in the importing ontology.

This allows an ontology to include classes from an external ontology without importing its axioms (this functionality is possible, but not frequently used) and thus its true definition through its intensionally defined semantic makeup. While this allows ontologies to reuse concepts, the inclusion of external classes without the inclusion of the axioms which define and restrict them means that potential incoherencies and inconsistencies

may arise between ontologies which reference each other. These could occur due to ignorance of the axioms which govern the class at the time of reference, or due to semantic shift in the source ontology over time. For example, to use the class *cell* (GO:0005623) in an ontology with the MIREOT method, it would suffice to use the class IRI from GO (i.e., `http://purl.obolibrary.org/obo/GO_0005623`), and add the axiom that *cell* (GO:0005623) is a subclass of *cellular component* (GO:0005575). MIREOT also allows importing some additional axioms from the referenced ontology, and recursive inclusion of parent classes. These features, however, are not always used. Axioms particularly are often discluded, to limit the amount of additional classes imported. Tools such as OntoFox facilitate MIREOT-ing terms from other ontologies, providing the necessary information in RDF/XML format for inclusion in an ontology. Another tool, Slimmer[65], constructs a new ontology by extracting and linking classes from a series of other relevant ontologies, as defined in a specification file.

More recently, several ontologies have been developed for specific applications that primarily consist of classes that are imported from other ontologies, and are combined in a new way suitable for the intended application, known as application ontologies. The Experimental Factor Ontology (EFO)[100] references classes from more than 26 other ontologies[142]. Originally developed to annotate data from gene expression experiments in the ArrayExpress and Gene Expression Atlas databases[131, 122], it is now applied to several additional domains, such as the annotation of disease to phenotype mappings in literature[134]. Another application ontology, eNanoMapper (ENM), was constructed by extracting classes relevant to the nanomaterials domain from many other ontologies, using the Slimmer tool[65].

The use of MIREOT, however, introduces the potential for ontology interoperability problems due to the exclusion or addition of axioms to ontologies between uses. Particularly, inconsistencies between class definitions may lead to unresolvable differences in conceptualisation, in turn leading to unsatisfiable classes or inconsistent ontologies when combined. This may be a problem both in the acute sense, that developers may acciden-

tally build upon another concept in a contradictory way, and in the sense that subsequent versions of the source ontology may re-axiomatise the subject class in a way which renders its use in the sourcing ontology incompatible with it. These problems may lead to unsatisfiable classes, and incoherent or inconsistent ontologies, when actually attempting to use them in combination. An earlier investigation into the Experimental Factor Ontology showed that the unchecked use of MIREOT had caused wide-spanning unsatisfiability and inconsistency[142]. However, while it has been shown that application ontologies contain interoperability problems with respect to their upper level ontologies, there is no analysis of interoperability between the upper level ontologies themselves, nor a solution for the resolution of inconsistencies.

Another issue with the MIREOT system is that ontologies which use it do not always include the required IRI of the source ontology when referencing a class, and thus it is difficult both to identify that a class is an imported MIREOT class, and to discover the authoritative definition for a class - especially if it has been heavily reused via MIREOT among the ontology corpus. For example, if we examine the *chi square test* class in the Clusters of Orthologous Groups ontology (CAO)[94], we notice its IRI is *http://purl.obolibrary.org/obo/OBI_0200200*. However, there is no assertion that identifies which ontology the class is referenced from. While a human can easily recognize that this IRI belongs to OBI[22], there is no provision for a computational reader to do this. This class is referenced in several ontologies: CAO, OBCS[176], OBI, and STATO[12]. This presents a challenge in tracking imported ontologies back to their source, which is necessary for the evaluation of maintained consistency after the use of MIREOT to import classes.

Ontology Mapping and Alignment

While ontology re-use is encoruaged, either through full ontology inclusion or via MIREOT, different ontologies may define terms that re-define concepts described in other ontologies. Part of the reason for this is that ontologies describe different domains of interest, and

The same is true of the *hypertension* (MP:0000231) exmaple given earlier: the HP class defines the class in the context of the human phenome, and the MP class of the mice phenome.

However, different ontologies may also define the same entities in the same context. This is particularly true of medical terminologies like SNOMED and ICD. The reason that these separate terminologies exist is political and historical. For example, the use of SNOMED requires a licence to be purchased on a government level, and countries which do not purchase the licence must use an alternative medical terminology, such as ICD. Legacy reasons are also a cause, such as in the case of MIMIC-II, which uses ICD-9 for diagnosis annotation, despite not being the current version of ICD[80].

The effect of this is that annotations of datasets may concern many of the same concepts of interest, but refer to different ontology terms. For example, in one hospital a patient could be annotated with the ICD-10 term *hypertension* (I10), while they would be annotated with *hypertension* (DOID:10763) in another hospital. Despite both patients suffering the same disease, it is not immediately obvious, especially to a computational observer, that they do. In fact, *hypertension* can be found in HP, SNOMED, ICD-9, ICD-10, and DOID, and other terminologies. The value of integrating these annotations is the value of having more data, and more data increases the power of data mining and analysis tasks. To do this, ontology mapping and alignment must be performed.

We previously discussed one method of integrating phenotypes across ontologies: via their logical formulations. Projects such as PheonmeNET attempt to use the axioms in phenotype ontologies, such as HP and MP, to create a new ontology which semantically links the relevant phenotypes through anatomy and qualities. However, many ontologies do not contain logical formulations, especially linking to the same external ontologies and axiomatised in the same way.

Official mappings also exist between certain terminologies. These can either be internal or external. In OWL ontologies, database cross-references are often used for this purpose. For example, the *hypertension* (DOID:10763) class contains cross-references to

(EFO:0000537), (ICD10CM:I10), (ICD9CM:401-405.99), and more. External mapping files are also published that describe links between terminology concepts. The NIH maintains an official mapping between ICD-9 and ICD-10, for example[6]. While manually created mappings cover common terms, they are limited in coverage. Some projects, such as MONDO, expand upon manual mappings by connecting concepts through intermediate database mappings. BioPortal also maintains a database of mappings sourced from different databases, which can be automatically queried and traversed via its API. Ontology alignment is also a major area of research, and is the subject of an annual event for method development and evaluation[8]. Tools such as AgreementMaker have been created to map ontologies on lexical and structural bases[41].

There are, however, problems that both automatic and manual mapping approaches cannot easily overcome. It is a design choice for ICD and other medical terminologies to pre-compose terms, while many other ontologies and terminologies prefer to post-compose them. To use the example of the obsoleted term earlier, *congenital neutropenia* (HP:0005549) was replaced, with instructions to annotate data using both *neutropenia* (HP:0001875) and the modifier *congenital onset* (HP:0003577). This is pre-composition: providing a minimal set of constituent terms in the terminology itself, then combining these in the annotations to describe more complicated objects. ICD-10, on the other hand, uses the term *congenital agranulocytosis* (D70.0) to cover the concept of congenital neutropenia. To map this term to HPO would require two mappings. Post-composed terms can also express even more complicated concepts, that would take three or more concepts from several different ontologies to represent, such as *Prolonged stay in weightless environment, occurrence on farm* (READ:YMBaG) from the Read Codes terminology[27]. The necessity for one-to-many and many-to-many mappings between biomedical terminologies compounds the problems of complexity and volume affecting both manual and automatic mapping methods. These difficulties are also part of the reason that legacy coding systems continue to be used, and instance annotation environments haven't been integrated.

Ontology Building and Access

Visual ontology construction can be performed using software such as Protégé, OBO-Edit, and WebProtégé[44, 117, 164]. The previously mentioned Slimmer tool is also a method of building ontologies using components from other ontologies. They can also be built with scripting languages and libraries, such as OWLAPI for JVM languages[75].

Ontology repositories are web-based applications that enable access to ontology features for both human and computational interaction. Most include both a web component and a computational API. They also allow interaction with multiple ontologies at once, such as via browsing database cross-references, and searching labels across many ontologies. Most ontology repositories only mediate access to the asserted or pre-inferred contents of ontologies. The OntoQuery software also provides reasoner-based ontology access over the Web[165], but limits itself to a small amount of ontologies per instance, though reasoner performance was found to be similar to that of working with tools such as Protégé.

AberOWL is a reasoner-based ontology access and analysis framework, which allows users to work with the semantic features of ontologies[73], its axioms and formal definitions, without the overhead of local classification and querying with a reasoner. It enables ontology-based analysis both with tools provided through its interface, and by facilitating ontology-based data access. It provides a web interface and an API which allow users to browse and explore ontologies through a reasoner. It overcomes the previously discussed difficulty of large-scale OWL reasoning by using the Elk reasoner[86], which only supports the EL subset of OWL, ignoring any axioms falllng outside it. This guarantees reasoning in polynomial time. Each ontology hosted by AberOWL is classified by Elk upon startup, and querying is performed by converting a Manchester OWL Syntax query into an OWL class expression using the OWL API[75]. Results may be returned from single ontologies or from the entire set of ontologies which AberOWL hosts.

2.2 Natural Language Processing

Finite state automata may be the philosophical and practical root of the field of natural language processing (NLP) when considered as a sub-field of computer science. However, it can also be considered as a sub-field of linguistics and information theory. In computer science, NLP is the process of extracting structured information from text. In Jurafsky's Speech and Natural Language Processing, the modern tradition of computational NLP is loosely split into 'four paradigms,' that span between 1970-2000+, each associated with characteristic algorithms that were developed or popularised during the period[82].

- 1970–1983
  - Stochastic modelling
  - Hidden Markov Models
  - Discourse Modelling
  - Reference Resolution
- 1983–1994
  - Finite State models
- 1994–1999
  - Incorporation of probability into previous methods and models
- 2000+
  - Machine Learning
  - Training on public datasets, with syntactic, semantic, and pragmatic annotations.

While machine learning remains popular in NLP, methods that were invented much earlier continue to be used and improved. Previously unfeasible models and methods can be made realistic with increasing computational power, combination with machine learning methods, or through the availability of public datasets on which to train. Just

as there are a multitide of methods and approaches to NLP, there are also many sub-tasks. Of particular interest to us are those used for text mining. Text mining is the task of extracting structured information from text. In a review of the field, Allahyari et al. list several text mining sub-tasks, including Information Retrieval (IR), Information Extraction (IE), sentiment analysis, and text summarisation[16]. In this thesis, we will focus on information extraction: automatically extracting structured information from text[18]. For example, the task of determining from a text document whether a particular patient has hypertension, or determining from a piece of literature whether hypertension is a symptom of heart failure. It is contrasted from IR, which is the task of finding relevant documents given a set of criteria. For example, finding documents concerning patients with heart failure.

IE tasks are usually implemented in the form of a series of components solving NLP tasks along a pipeline towards the intended result. Allahyari et al. further delineate two major tasks in IE: named entity recognition (NER, and also referred to as annotation) and relation extraction. NER is the extraction of named entities of interest from a text document (e.g. recognising a mention of hypertension in a clinical narrative)[16, 114]. The relation extraction task builds on the results of the NER task, to identify relationships between the entities involved. For example, whether a patient suffers from hypertension or not. These tasks, in combination, consistute information extraction.

Before either NER or relation extraction take place, text is pre-processed. This involves word-level transformations such stemming and lemmatisation, which obtain the root of the word and the word in its uninflected form respectively. Uninformative or common words are also removed (frequently called stop words). Words can then be transformed into tokens, with multiple words describing single entities combined into single tokens, and certain uninformative punctuation thrown away.

There are many frameworks for NLP and information extraction, which span many different approaches. General frameworks for NLP such as Stanford CoreNLP[102], nltk[96], and GATE[42] include implementations of many of the sub-tasks involved in information

extraction pipelines, such as NER. The relation extraction process, however, is usually performed on an individual basis, or as part of an integrated pipeline. For medical data, integrated pipelines such as CogStack combine the features of these lower level information extraction libraries and extend them with tools for relation extraction, manual validation, and querying[78].

2.2.1 Negation Detection

A major component of information extraction pipelines are algorithms that determine the context of an entity. It is only with information concerning the context in which an entity has been mentioned in text that the overall relationship between an object and subject can be determined. This is a critical part of the relation extraction process. For example, a mention of a disease in a clinical letter does not mean that a patient suffers that disease; many clinical letters discuss a diagnostic process with respect to a disease. A letter may, therefore, discuss a test being conducted to determine whether a patient has a condition. The letter may also rule out the condition, or only mention that it is present in a family member. There may also be a level of uncertainty expressed as to whether a patient actually has a condition.

We will focus on negation detection algorithms. In the context of biomedicine, negation detection algorithms determine whether a finding mentioned in a clinical text narrative is stated as absent or present, usually using the sentence mentioning the concept as input [114]. Algorithms for negation detection can be split between rule-based and machine-learning approaches. In this thesis, we will consider rule-based approaches, because they are explainable. To develop a tool that allows clinicians to validate and trust decision procedures, the reasons for the decision need to be clear. This is both a practical and an ethical concern. Furthermore, the performance of rule-based methods for negation in comparison to machine-learning methods is unclear, but favours rule-based approaches in the literature. Goryachev and Taggart et al. compared rule-based and trained ML approaches directly, and found rule based approaches to outperform ML approaches in

the considered contexts[59, 155]. Wu et al. found that ML models modestly outperformed rule-based classifiers with out-of-context training, yielding further improvements with in-context training[173]. However, this work compares several machine learning classifiers with one particular implementation of NegEx that is not used in any of the previous three studies that found superior rule-based performance, and does not consider it for additional training in any context. Another ML negation work, Taylor and Harabagiu evaluated several machine learning methods, but did not consider any rule-based methods, even though the performance was similar to rule-based methods presented elsewhere in the literature[158].

A popular rule-based approach to negation detection for clinical documents is NegEx, which uses regular expressions to determine the negation of a concept in a sentence[32]. This was later generalised into ConText, which remains in frequent use today[64]. For example, it is used for negation detection by CogStack[78]. Later work has built on ConText, extending it using graph-based algorithms that determine negation through typed dependency relationships generated by a dependency resolution task. DEEPEN[110] works in this way, operating only upon concepts that NegEx determines to be negated. Other dependency-based algorithms make no use of NegEx, such as NegBio, negation-detection, and DepNeg[125, 55, 147].

Instead of specific lexical patterns, dependency-based approaches define grammatical patterns to determine whether or not a concept is negated in a sentence. The hypothesis is that grammatical patterns are more generalisable, and discernment is attuned to grammatical nuance beyond the mere mention of a word, and training should therefore transfer better to internal test sets, as well as to external data. As reported in the papers that present these algorithms, dependency-based algorithms show an improved precision over syntactic approaches. However, with the exception of negation-detection, they require the development of specific grammatical rules, which is time-consuming and does not generalise well. ConText rules are ignorant of grammatical relationships, and a comparison showed that ConText maintains performance on a new datasets while others

do not[59]. Approaches such as SynNeg have attempted to extend ConText with more specific grammatical rules, but showed only modest performance improvements[157].

2.2.2 Text Mining and Ontologies

As discussed previously, metadata are a common feature of biomedical ontologies. These metadata include a wealth of natural language information, and this makes them a valuable resource for text mining[148]. Particularly, ontology classes are often associated with labels, which can act as vocabularies for information retrieval and named entity recognition tasks.

Associating text with ontology terms is known as semantic annotation. Semantic annotation enables integration with any other data annotated using the ontology, including in other modalities (such as image data or structured data). It also enables integration with the ontology structure itself, and this can be useful both during and following annotation. The structure of the ontology can be used to automatically mine for concepts. For example, in an information extraction experiment which aims to search literature for articles concerning monogenic diseases, a vocabulary could be built using the subclasses of *monogenic disease* (DOID:0060340) in the Human Disease Ontology (DO)[136]. This would, at the time of writing, yield 2,634 classes, and more than 7,552 labels and synonyms. In addition to gaining labels for a large number of diseases, the annotation vocabulary also provides additional power via the intermediate classes returned. For example, one of the subclasses of monogenic disease is *autosomal genetic disease* (DOID:0050739); not only would all articles concerning the individual autosomal genetic diseases be returned, but also any articles mentioning this more general concept. Web-based ontology annotation services, such as the NCBO Annotator and Skylark, support annotation of groups of ontology terms based on structural queries. The AberOWL ontology repository, in addition, supports annotation of PubMed articles with classes satisfying a complex description expressed by description logic formulae (support has been removed in subsequent versions).

Following annotation, the structure of the ontology can be leveraged for ontology-based

analysis. Techniques such as semantic similarity, semantic rule-mining, or relational machine learning can be used as an analytic method for entities described by text. This can be an alternative to methods such as word embeddings or recurrent neural networks, which can use text as a training examples for artificial intelligence tasks. Some methods, such as Onto2Vec, learn word embeddings directly from ontology axioms and annotations[51].

Vocabulary Expansion

OWL ontologies define annotation properties that can be used to describe multiple natural language labels for a single concept. Open Biomedical Ontologies[160] define a series of conventional annotation properties that can be used for the expression of labels and synonyms in biomedical ontologies. These features are widely used in biomedical ontologies; as of 2017 the Human Phenotype Ontology (HP)[87] contained 14,328 synonyms for 11,813 classes[89]. Because such labels are associated with ontology terms, ontologies constitute a controlled domain vocabulary. Their provision of vocabularies, and the standard use of ontologies for data annotation, makes them an important resource for information retrieval and named entity recognition tasks[148].

However, due to limitations in resources for expert curation of ontologies and the sheer scale of their contents, synonym lists and labels obtainable from ontologies are not exhaustive. Combined with the tendency for alternative descriptions of semantically equivalent concepts in biomedical text[35], ontology labels are not always a good fit for text corpora that discuss the same concepts[30]. By expanding the set of synonyms in an ontology, particularly with synonyms that provide a better fit for text corpora, the performance of natural language processing tasks that depend on them is necessarily improved.

This potential is reflected by previous work in the field. One approach that used analysis of existing synonyms across ontology hierarchy to determine new synonyms reported an increase in performance of a task retrieving articles from a literature repository[154]. Another rule-based synonym expansion approach to extending the Gene Ontology showed

improved performance in named entity recognition tasks[50]. A combined machine-learning and rule-based approach to learning new HP synonyms from manually annotated PubMed abstracts improved performance of an annotation task over a gold standard text corpus[95].

Outside of automated synonym generation, organised efforts have been made to manually extend an ontology's synonyms for a particular purpose. For example, HP was expanded with layperson synonyms to enable its use in applications that interact directly with patients[90].

Ontology-based annotation software such as OBO Annotator[60], ConceptMapper[60], and the NCBO Annotator[81] contain routines to consider rule-based morphological and positional transformations of terms to increase NER recall. Parameters that control the use of these features have a strong influence on annotation performance[49]. Previous work has also investigated synonym acquisition and derivation for the purposes of improving the performance of lexical ontology matching and alignment tasks[126].

2.2.3 Classification and Survival Analysis

Classification and survival analysis are major tasks in statistics and machine learning. They are both kinds of supervised learning task, wherein examples from labelled data are used to create a model. In classification, the model attempts to categorise new observations, while in survival analysis, covariates are related to the amount of time that passes before a particular event occurs. In this thesis, we are interested in these tasks in the context of diagnostic and prognostic modelling. Classification is the task used for diagnostic modelling, which is prediction of the likelihood of a particular diagnosis or label. Survival analysis is used for prognostic modeling, which is the prediction of the risk of a particular outcome during a time period of interest. Though the research designs differ, similar methods and frameworks are used for both. Henceforth, we will use prediction models to refer to both tasks.

As well as predicting the likelihood of a diagnosis or outcome for unseen individuals,

prediction models can identify independent risk factors. For example, a large review of prediction models by the Stroke Risk in Atrial Fibrillation Working Group found that prior stroke, advancing age, hypertension, and diabetes are reliable independent risk factors for stroke in atrial fibrillation patients[152]. Identification of risk factors for disease and outcomes can be used to improve outcomes in several ways. They can inform public health information, as in the case of warnings provided on cigarette packs[33]. They also directly inform clinical practice. The European Society of Cardiology (ESC) and European Society of Hypertension (ESH) jointly publish guidelines for the management of arterial hypertension in clinical practice, uses information about risk factors for cardiovascular disease and other diseases to suggest preventative and mediative treatments, such as in the case of comorbid diabetes[172]. Prediction models can also be used to inform advice against the use of traditional or folk treatments, such as aspirin for hypertension[43].

As well as informing guidelines, prediction models can also be used to construct tools that are used directly in clinical practice. In the management of atrial fibrillation, the CHA_2DS_2-VASc and HAS-BLED scores are used for decision making to support the prevention of stroke[92]. Such tools are often provided in the form of a web-based calculator, and can also be integrated into the EHR system.

Many different technologies and methods, across several domains of research are used for predictive modelling. In this thesis, we will concentrate on linear regression models. A linear regression determines a linear relationship between a dependent variables, and a number of co-variates or predictors. Linear regression models are widely used for prediction modelling with healthcare data because of their relative simplicity, interpretability, and reproducibility. Regression models also produce beta-coefficients that describe likelihood ratios, and can be used as evidence (though not alone) for causal relationships.

Tu discussed the advantages and disadvantages of using artificial neural networks over regression models for medical outcomes, and found that while neural networks will more easily discover non-linear relationships without advanced statistical knowledge, it is much harder to actually recover any relationships between predictors and the outcome[163].

Although some work reports that machine learning models outperform linear regression models[139], a systematic review showed no overall improvement in performance[34]. Non-linear regression models can also be used to model non-linear relationships, but are not considered in this thesis.

Particularly, we will use Cox models[40]. Cox models are a special case of regression, for which the dependent variable is the hazard function at a given time point, rather than a probability of whether or not the event will happen. The Framingham CVD Risk Score is a widely used example of a model produced with a Cox regression.

Bellazzi and Zupan, in a discussion of predictive data mining, delineate the following tasks as a general pipeline for model development[23]:

1. Defining the problem, setting the goals
2. Data preparation
3. Modelling and evaluation
4. Construction of the target predictive model
5. Deployment and dissemination

Defining the problem involves determining the outcome of interest, and the methods to be used. This involves identifying the dataset, doing background research on the domain of interest. The data must then be acquired, and pre-processed. Depending on the validation model, it may be split into training and test sets at this point. Analysis of bias risk should also take place at this stage. Hayden et al. presented a QUality In Prognostic Studies (QUIPS) tool, which recommends six domains for bias consideration[66]:

1. Study participation
2. Study attrition
3. Prognostic factor measurement
4. Outcome measurement
5. Study confounding
6. Statistical analysis and reporting

As well as determining potential bias, these instruments also help the researcher to determine whether the kind of data collected is actually useful or appropriate for the purposes of the study. After this, the actual modelling is performed. In a discussion of practical approaches to predictive model development, Steyerberg describes seven steps for model development:

1. Data inspection
2. Specification
3. Estimation
4. Performance
5. Validation
6. Presentation

These steps encompass the modelling, evaluation, and target model construction steps described by Bellazzi and Zupan. Data inspection involves some of the tasks mentioned previously, especially around determining potential bias. Initial explorations using unsupervised methods such as clustering and descriptive statistics may help to identify outcomes of interest or hypotheses for research questions. It also includes analysis of data missingness, dichotomising continuous predictors, and decisions on which predictors to consider in the model.

Continuous predictors are often combined into dichotomous factors, because they are more easily interpretable, and their effects on survival can easily be visualised as kaplan meier curves. This can either be done based on background knowledge (e.g. categories of BMI), equal frequency binning, or with respect to the outcome variable. Despite the advantages in interpretability, categorising predictors necessarily damages resolution[17]. In practice, categorisation of variables has been shown to produce predictive models with "poor predictive performance and poor clinical usefulness[17]."

Specification involves feature and model selection. This is often based on data available, and limits to the amount of events that can be used without overfitting. This is usually framed in terms of Events Per Variables (EPV). Ten events per variables has been

a long-time standard rule of thumb for prognostic models[124], but various simulation studies have suggested different values as low as 5 EPV[124]. The number of events per variable includes variables considered by any feature selection in the model itself, and therefore pre-selection of features often takes place based on expert knowledge, or literature review. This limits the ability for models to discover new relationships, and certain approaches such as Lasso have been used for feature selection[177]. Pre-selection can also involve identifying highly correlated variables, since groups of highly correlated variables will not be informative.

Model assumptions should also be tested. For example, in a proportional hazards model, hazard proportions should remain constant throughout the period of interest. This may not be the case for certain predictors; for example, a certain blood measurement may represent a large short term risk, but a low long term risk. If this is the case, then a multiplactive value applied to the increasing value of a covariate (a hazard ratio) does not make sense, becuase it cannot capture both relationships.

In the case of categorical variables, this can be done by examining the kaplan meier curve for any crossover or drop-off. For continuous variables it is more complicated, but schoenfield residuals can be assessed either visually or by summative statistics to determine whether the variable has a relationship with time. In the cases of small or insignificant time-dependent relationships, the effect can be ignored as long as it is noted in the methodology as a potential source of bias. Otherwise, variables can be split into time-dependent variables. Non-linear relationships can be transformed using logarithms or fractional polynomials. Transformations can either be chosen manually, or determined automatically with decision processes that measure the variable's relations

The model will be fitted at this point, and initial performance can be reported. This is known as the estimate. Feature selection can be performed inside the model by performed in several ways, usually via backwards or frontwards stepwise elimination, which remove uninformative variables while attempting to preserve any information gained from relationships.

Prediction models are evaluated in terms of their discrimination and calibration[151]. Discrimination is a measure of how well a predictive model can discern between patients with or without a particular target outcome, while calibration is a measure of agreement between observed and predicted risk. Calibration is also a good indicator of how much the model is overfitting. Due to the tendency of predictive models to overfit, it is also recommended that studies report the estimated performance metrics adjusted for optimism, as achieved by bootstrapping or a similar method.

The presentation of a model is an important component for scientific value. It is vital that enough information about the study design, results, and validation is shared to allow readers of the study to discern whether or not the study was well designed, and whether the model performed well. Models must also be reconstructable if they are to be used or externally validated. TRIPOD (Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis) is a set of reporting guidelines for prediction models[36]. It includes a checklist for model developers, to ensure that they report the model they develop in such a way that they are reproducible. Several studies have found, however, that a majority of prediction model publications do not meet these standards[68, 25]. This problem also exists for machine learning models, because they cannot easily be shared or re-implemented for external validation. Such studies warn against the over-reporting of results, and TRIPOD notes that model reporting should include discussions of model limitations.

2.3 Hypertrophic Cardiomyopathy

Hypertrophic Cardiomyopathy is a common inherited disease defined by otherwise unexplained thickening of the heart muscle, whose first modern description appeared in the literature by in 1958[159]. First thought to be rare and untreatable, it is now known to be a common disease, with a prevalence of around one in five hundred people[103], which is highly treatable and whose patients mostly maintain their quality and longevity

of life[106]. Nevertheless, it has a strong public profile, because it is the most frequent cause of sudden death in young adults[104], including elite atheletes[105], and can lead to heart failure or stroke.

The major challenge of HCM is its heterogenous and complex presentation, genotype, and phenotype[171]. It is difficult to diagnose, treat, and manage, particularly for non-specialists in the area[107]. The identification and treatment of HCM falls under two primary schools of thought, one produced by the European Society of Cardiology (ESC), and the American College of Cardiology Foundation/American Heart Association (ACCF/AHA), and are known as the ESC guidelines[13] and ACCF/AHA guidelines[53] respectively. In 2003, both organisations published a joint clinical consensus, although both groups have since produced updates their own seperate guidelines.

Udelson James E. discussed a "trans-atlantic divergence" between the two guidelines' approaches for reducing the risk of sudden death[166].

Since most patients with HCM do not suffer sudden death, and because it is the least manageable of its complications, most predictive models have focused on prediction of sudden death in HCM. These patients can then be prophylactically fitted with implantable cardioverters/defibrillators (ICDs). There are two widely used risk scoring systems derived from time-to-event prognostic models, the Enhanced American College of Cardiology/American Heart Association Strategy (ACC/AHA), and the HCM risk-SCD score[120].

However, sudden cardiac death is not the only cause of death in HCM; it can also lead to cardiovascular death via the development of heart failure and stroke[108]. These causes of mortality are included in a wider set of complications of the disease. Cardiomyopathy UK describes the following complications of HCM[5] (descriptions quoted from relevant ontologies).

Arrythmias "Any cardiac rhythm other than the normal sinus rhythm. Such a rhythm may be either of sinus or ectopic origin and either regular or irregular. An arrhythmia may be due to a disturbance in impulse formation or conduction or both[87]."

(HP:0011675)

Atrial fibrillation "An atrial arrhythmia characterized by disorganized atrial activity without discrete P waves on the surface EKG, but instead by an undulating baseline or more sharply circumscribed atrial deflections of varying amplitude an frequency ranging from 350 to 600 per minute[87]." (HP:0005110)

Ventricular tachycardia "A tachycardia originating in the ventricles characterized by rapid heart rate (over 100 beats per minute) and broad QRS complexes (over 120 ms).[87]" (HP:004756)

Ventricular fibrillation "Uncontrolled contractions of muscles fibers in the left ventricle not producing contraction of the left ventricle. Ventricular fibrillation usually begins with a ventricular premature contraction and a short run of rapid ventricular tachycardia degenerating into uncoordinating ventricular fibrillations.[87]" (HP:0001663)

Heart Failure "The presence of an abnormality of cardiac function that is responsible for the failure of the heart to pump blood at a rate that is commensurate with the needs of the tissues or a state in which abnormally elevated filling pressures are required for the heart to do so. Heart failure is frequently related to a defect in myocardial contraction[87]." (HP:001635)

Stroke "Sudden impairment of blood flow to a part of the brain due to occlusion or rupture of an artery to the brain[87]." (HP:0001297)

Sudden cardiac death "The heart suddenly and unexpectedly stops beating resulting in death within a short time period (generally within 1 h of symptom onset)[87]." (HP:0001645)

Olivotto et al. explores the relationship between AF and HCM, finding a 22 percent incidence of AF among HCM patients, and that these patients have a significantly greater chance of HCM-related death due to heart failure[119]. They also showed a higher

probability of combined HCM-related death, functional impairment, and stroke. Siontis Konstantinos C. et al. also find a strong relationship between AF and mortality in HCM. Meanwhile, heart failure itself is a leading cause of disability and death, with annual mortality reported as high as 17.5%[168]. One study into heart failure in HCM found that 17% of patients developed heart feailure, with 20% of those patients dying[111].

CHAPTER 3

UNMIREOT

3.1 Introduction

In the literature review we showed that the MIREOT system for referencing external ontology terms enables large scale ontology integration and re-use. We also discussed that the practice can lead to issues of interoperability that are hidden from automated reasoners when considering MIREOT-ed classes without the context provided by the axioms defined in the ontology being referenced. Particularly, we discussed a piece of previous work, wherein problems of hidden unsatisfiability were revealed in the Experimental Factors Ontology (EFO) when it was combined with the ontologies it references[142].

The extent to which this problem is shared by the rest of the biomedical ontology ecosystem, however, is unknown. It is also unknown whether there are common roots to any widespread unsatisfiability, and whether the unsatisfiability can be automatically repaired. In this chapter, we provide a case-study on interoperability and unsatisfiability between a core set of biomedical ontologies that encourage term re-use between them: the OBO Foundry[160].

To achieve this, we extend the tool described by Slater et al. for the analysis of interoperability in EFO, generalising it to reveal hidden contradictions in any combination of OWL ontologies[142]. We then use the tool to evaluate the OBO Foundry for hidden cases of inconsistency and unsatisfiability, reporting upon the sources of inconsistency and

the most implicated axioms.

We then present a novel algorithm that uses unsatisfiability explanations to repair all cases of hidden inconsistency in an ontology by iteratively removing the most implicated axioms. The algorithm uses iterative random sampling of maximally general classes to repair even very large ontologies with tens of thousands of unsatisfiable classes in a few iterations. This approach works because unsatisfiability is transitively inheritable through the subclass relation. For example, if a 'disease' class were unsatisfiable, so would all subclasses of disease specifying it be unsatisfiable. Moreover, if the cause of unsatisfiability for the disease class was repaired, it would also repair that inherited cause of unsatisfiability for all its subclasses. By grouping high level unsatisfiable terms, and removing their most frequently implicated axioms, dependent groups of unsatisfiabile terms are repaired without having to examine every class.

3.2 Materials and Methods

All non-deprecated and obtainable OBO ontologies were downloaded using the permanent download links given by the OBO Foundry database at `http://obofoundry.org/registry/ontologies.yml`. They were obtained on 28/03/2018, for a total of 132 ontologies.

We first worked with the OBO Foundry Core ontologies, listed in Figure 3.1 (excluding the Protein Ontology (PRO), as it was unobtainable). These ontologies are judged as satisfying the OBO Foundry principles, and are therefore tightly integrated, and are heavily used throughout the rest of the OBO Foundry. We combined these ontologies with all of the ontologies in their import closures. We then identified transitive ontology imports, particularly finding that the Plant Ontology (PO) imports NCBI Taxon, while DO includes another 8 ontologies. Since the combined consistency of these ontologies is evaluated later in the experiment, we did not combine them into the OBO Foundry core. Subsequently, we evaluated this combined meta-ontology for cases of unsatisfiability and

their etiology.

BFO Basic Formal Ontology[145]

CHEBI Chemical Entities of Biological Interest[45]

DO Disease Ontology[136]

GO Gene Ontology[20]

OBI Ontology for Biomedical Investigations[22]

PATO Phenotypic Quality Ontology[56]

PO Plant Ontology[79]

XAO Xenopus Anatomy and Development Ontology[137]

ZFA Zebrafish Anatomy and Development Ontology[149]

Figure 3.1: Ontologies included in the OBO Foundry Core.

We then developed an algorithm for quickly identifying a small set of axioms that can be removed from an ontology to solve all cases of unsatisfiability, using an understanding of Reiter's theory of system diagnosis, which has previously been used to develop algorithms for unsatisfiability justification in the domain, and is also the basis of the naive HST algorithm [85]. We apply this algorithm to the combined OBO Foundry Core meta-ontology, creating a coherent version of it. We combine this coherent core variant iteratively with every other ontology in the OBO Foundry individually, counting the unsatisfiable classes revealed in each case. Following this, we apply the unsatifisability repair algorithm to every ontology-core combination found to contain unsatisfiable classes, noting the axioms removed and counting them. Using these results, we report on the most widely implicated axioms causing unsatisfiability across the OBO Foundry, and evaluate some of the root causes.

3.2.1 Implementation and Experimental Setup

For all experiments, we use the OWLAPI 5.1.4[75]. To classify the ontologies and to retrieve unsatisfiability explanations, we use the Elk reasoner version 0.5.0-SNAPSHOT[86].

Elk supports the OWL 2 EL profile, a fragment of OWL that supports tractable (i.e., polynomial-time) reasoning, but which lacks support for many logic operators. In particular, OWL 2 EL does not support the use of negation in class descriptions or use of the universal quantifier. The only type of axiom in OWL 2 EL that could result in an explicit contradiction is the disjointness axiom. We also used Protégé to examine some of the combined ontologies for particular cases of unsatisfiability[118].

3.3 Results

All tools described in this chapter, including those to obtain, merge, analyse, and repair ontologies, are available at `https://github.com/bio-ontology-research-group/UNMIREOT`.

3.3.1 Combining ontologies and detecting inconsistencies

The 9 core ontologies combined consist of 402,868 logical axioms and 207,105 class declarations, of which 636 were unsatisfiable. Table 3.1 shows the distribution of unsatisfiable classes. The source of the classes were determined using the IRI prefix, which are defined for each ontology in the OBO Foundry metadata file.

Table 3.1: Unsatisfiable class counts in OBO Foundry

Ontology	Unsatisfiable Class Count
CHEBI	37
GO	565
OBI	34

Upon combining the repaired version of the OBO Foundry core with each of the 130 ontologies in the OBO Foundry, we revealed unsatisfiable classes in 46 of them. The

ten OBO Foundry ontologies with the most hidden unsatisfiable classes are listed in Table 3.2. The total number of unsatisfiable classes was 343,381, while the total number of unique unsatisfiable classes was 204,033. Of these, 8,891 were obsolete classes (and are intentionally unsatisfiable).

Table 3.2: The ten ontologies with the most unsatisfiable classes in the OBO Foundry, when combined with a repaired version of the merged core ontologies.

Ontology Name	**Unsatisfiable Class Count**
MONDO[138]	97,340
UPHENO[88]	88,479
OMIT[76]	63,015
MOP[128]	57,355
RXNO[132]	57,330
HP[87]	46,031
MP[146]	43,762
OBA[48]	26,523
OAE[67]	20,566
NBO[57]	20,038

3.3.2 Ranking and repairing axioms

As described above, our investigations created extremely large combined ontologies, with a great number of unsatisfiable classes. In the most prolific case, the MONARCH Disease Ontology (MONDO) contained just short of 100,000 unsatisfiable classes. Reiter's Hitting Set Tree (HST) algorithm can be used to generate all explanations for any number of unsatisfiable classes in an ontology. However, this algorithm runs exponential to the size of the conflicting sets of classes considered, and does not have a practical upper bound on running time[98].

Reiter's general theory of system diagnosis has been used in discussions of unsatisfiability justification for OWL ontologies, and also forms the basis of the naive HST algorithm. It considers a series of conflict sets, which describe conflicting sets of components in a malfunctioning system. In each conflict set, at least one component must be removed to repair the conflict. A hitting set is a set of components that intersects every conflict set,

and the hitting set problem is the task of computing all the minimally sized hitting sets for all conflict sets in the system, and therefore the minimal sets of components that can be removed to repair all malfunctioning components in the system [130]. Our problem can be reduced to the hitting set problem, as we have a set of unsatisfiable classes, which each have an unsatisfiability justification: a list of axioms, of which any one can be removed to repair that case of unsatisfiability. To remove all cases of unsatisfiability, a hitting set of axioms that intersect all unsatisfiable classes must be removed from the ontology. Since we cannot exhaustively generate justifications for all unsatisfiable classes, we developed an approach that makes use of the taxonomic structure of OWL ontologies to minimise the number of justifications.

We created an algorithm that iteratively samples high-level ontology classes, whose most implicated axioms can then be removed, solving causes of unsatisfiability for their subclasses without directly querying them for justifications. The algorithm is shown in Figure 3.2. It finds the group of unsatisfiable classes with the most asserted subclasses in the ontology, and does not have an unsatisfiable superclass. This ensures that it finds groups of general terms whose reasons for unsatisfiability account for a maximal number of additional terms. We use direct, asserted axioms instead of transitively inferred axioms because, as unsatisfiable classes, their inferences may be wrong. The effects of a more specific class potentially having more direct subclasses than the more general class is controlled by removal from consideration any classes that have an unsatisfiable superclass in the set. By solving for groups of general axioms, instead of just one, we make it more likely to find axioms that account for unsatisfiability shared across several pathways of subclass inheritance. By iteratively performing this process, we solve all cases of unsatisfiability by solving for groups of high level cases of unsatisfiability until there are none left, leading to the identification and removal of a minimal set of maximally implicated axioms.

Throughout execution of the algorithm, classes repaired by the removal of each axiom are recorded, and then counted and summarised at the end. This enables its use by

```
Data: o = Given ontology
Result: Minimal set of axioms required to repair ontology
Load and classify ontology o
x = unsatisfiable classes
while x > 0 do
    y = x without leaf classes (zero subclasses in o)
    if size(y) < size(x) then
        y = y without classes which have a superclass in set y
        z = group classes in y by total number of direct subclasses in o
        x = max(z.key)
    end
    if size(x) > 25 then
        x = randomly sample 25 classes from x
    end
    c = implicated axioms for each class in x
    Count axioms in c, and remove the maximally implicated axiom from o
    Reclassify ontology o
end
```

Figure 3.2: Algorithm for automatic repair of unsatisfiable classes in ontology.

ontology developers for identifying problematic axioms affecting groups of ontologies. It also enables us to identify problematic axioms causing unsatisfiability across many groups of ontologies.

3.3.3 Application to OBO Foundry

We applied the auto-repair algorithm first to the merged OBO Foundry core ontology, finding that two axioms could be removed to solve all cases of unsatisfiability.

1. *realizable entity* (BFO:0000017) SubClassOf *specifically dependent continuant* (BFO:0000020) with 599 implications.

2. *molecular entity* (CHEBI:23367) SubClassOf *material entity* (BFO:0000040) with 37 implications.

While this minimal set of two axioms were determined for removal based on maximal implications for class unsatisfiability, the actual causes of class unsatisfiability derive from violation of disjointness axioms, i.e. they are equivalent to or subclasses of two

or more classes asserted to be disjoint. In fact, the removal of one such axiom actually solves multiple disjointness violations. For the first and most prolific axiom, the classes it accounts for each violate one or more of these three different disjointness axioms:

1. *independent continuant* (BFO:0000004) DisjointWith *specifically dependent continuant*
   (BFO:0000020)

2. DisjointClasses: *independent continuant* (BFO:0000004), *specifically dependent continuant*
   (BFO:0000020), *generically dependent continuant* (BFO:0000031)

3. *continuant* (BFO:0000002) DisjointWith *occurrent* (BFO:0000003)

The second case is accounted for by two disjointness axioms:

1. *independent continuant* (BFO:0000004) DisjointWith *specifically dependent continuant* (BFO:0000020)

2. DisjointClasses: *independent continuant* (BFO:0000004), *specifically dependent continuant*
   (BFO:0000020), *generically dependent continuant* (BFO:0000031)

The two disjointness axioms shown for the second case are included in the three axioms shown for the first set. In total, therefore, three disjointness axioms account for all cases of hidden unsatisfiability throughout the OBO Foundry core ontology. Removing the subclass axioms is a more efficient and minimal route to solving the cases of unsatisfiability, because they prevent terms from violating multiple disjointness axioms. For example, in the case of removing the subclass relationship between *molecular entity* (CHEBI:22367) and *material entity* (BFO:0000040). Some subclasses of molecular entity violate the first disjointness axiom, and some violate the second. By removing the subclass axiom, however, molecular entities are no longer subclasses of material entity's parent *independent continuant* (BFO:0000004), which is a feature of both disjointness axioms.

Table 3.3: Top ten axioms accounting for the most hidden cases of unsatisfiability across OBO Foundry.

Axiom	Class Count
miRNA_target_gene_primary_transcript (NCRO:0000001) SubClassOf *nc_primary_transcript* (SO:0000483)	59,887
has role (RO:0000087) ObjectPropertyRange *role* (BFO:0000023)	57,335
processual entity (UBERON:0000000) SubClassOf *occurrent* (BFO:0000003)	41,675
processual entity (UBERON:0000000) DisjointWith *anatomical entity* (UBERON:0001062)	36,156
organ (UBERON:0000062) SubClassOf *has 2D boundary* (RO:0002002) ObjectSomeValuesFrom *anatomical surface* (UBERON:0006984)	19,797
disposition (BFO:0000016) SubClassOf *realizable entity* (BFO:0000017)	13,167
obsolete_disease (OBI:1110055) SubClassOf *ObsoleteClass* (GO:ObsoleteClass)	8,880
continuant (BFO:0000002) DisjointWith *occurrent* (BFO:0000003)	8,237
steroid hormone (CHEBI:26764) SubClassOf *steroid* (CHEBI:35341)	4,242
molecular entity (CHEBI:23367) SubClassOf *material entity* (BFO:0000040)	2,527

Among the wider set of OBO Foundry ontologies, we found that a set of only 55 axioms accounted for all 323,381 unsatisfiable classes. Of these, 28 involved a BFO class. Figure 3.3 shows the top ten axioms ranked by the number of unique unsatisfiable classes they are responsible for.

3.3.4 Inconsistency Analysis

The results of the algorithm show which axioms can be removed to solve cases of unsatisfiability, and further analysis reveals which disjointness axioms are most frequently violated. This alone, however, does not reveal the semantic misconception lying at the root unsatisfiability. While 599 hidden unsatisfiable classes were repaired in OBO Foundry core by removing the subclass axiom, *realizable entity* (BFO:0000017) SubClassOf *specifically dependent continuant* (BFO:0000020), this does not mean that the axiom, or the disjointness axioms it is associated with are themselves incorrect.

Instead, the errors occur through incorrect use of these terms by more specific classes. 87 of these 599 classes are *MAP kinase activity* (GO:0004707) and its subclasses. The

disjointness axiom they violate is the fundamental BFO distinction between *continuant* (BFO:0000002) and *occurrent* (BFO:0000003). A continuant is something that maintains its identity over time, while an occurrent is a temporal event. They are usually used in biomedical ontologies to distinguish between material entities and events or processes.

As shown in Figure 3.3, *MAP kinase activity* is a transitive subclass of *continuant* by means of being a *molecular function*. It is also a subclass of *part of* a *MAPK cascade*, which is a subclass of *intracellular signal transduction*. This class stands in an *occurs in* relationship with *intracellular*. The object property *occurs in* contains a restriction of its domain, asserting that something that *occurs in* something must be an *occurrent*.

So, *MAPK cascade*, a kind of *intracellular signal transduction*, is something that occurs *intracellular*ly. Because *MAP kinase activity* is *part of* a *MAPK cascade*, it is also an *occurrent*. The reason for this is that the *part of* (BFO:0000050) relationship must be between two things of the same kind; its definition states "two distinct things cannot be part of each other," and this restriction is inferred by the classifier from a *part of* relationship assertion. That *MAP kinase activity* must be both a *continuant* and an *occurrent* is the source of its unsatisfiability.

In addition to the 87 classes that are accounted for by *MAP kinase activity*, in fact all 599 unsatisfiable classes repaired by the *realizable entity* (BFO:0000017) SubClassOf *specifically dependent continuant* (BFO:0000020) axiom are subclasses of the class description:

- *molecular_function* AND *occurs in* SOME *intracellular*

This is fundamentally the same cause for unsatisfiability as *MAP kinase activity*: that they are subclasses of *continuant* via *molecular_function*, and *occurrent* via being something or a part of something that *occurs in intracellular*. There are actually 1,306 total classes which are subclasses of *occurs in some intracellular*, but 707 of these are not also subclasses of continuant, and are therefore not unsatisfiable.

These contradictions are not revealed by the automated reasoner used on the Gene Ontology alone, because it imports *occurs in* (BFO:0000066) from the Relation Ontol-

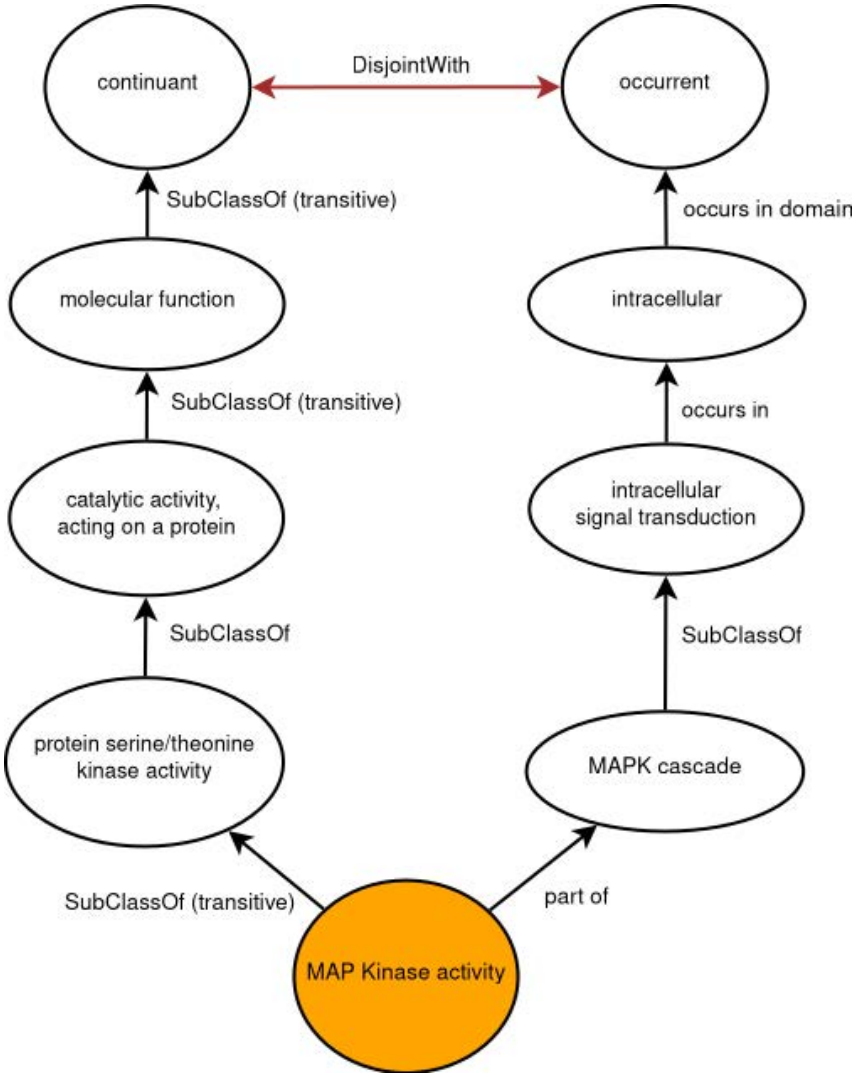

Figure 3.3: MAP Kinase unsatisfiability represented as a graph.

ogy via MIREOT, without its axioms. Particularly, the axiom that asserts its domain must be a *process* (BFO:0000015) (a kind of *occurrent*), while the range must be an *independent continuant* (BFO:0000004). The contradiction is revealed when the ontologies are combined and the imported class is therefore extended with its original axiomatic restrictions.

The mistake in modelling itself is easy to make. The shared inheritance of continuant and occurrent are hidden behind several layers of subclass and object property relationships. Furthermore, in colloquial language, we might easily describe a relationship between a part and a whole of different classes. The problems could be fixed without any destructive changes to the ontology by instead using the *participates in* (RO:0000056) or

has participant (RO:0000057) instead of *part of*.

3.4 Discussion

We have shown that there is a high prevalence of hidden inconsistency throughout a major biomedical ontology ecosystem, which includes widely used fundamental ontologies. We also presented an algorithm that can repair incoherent ontologies by removing a small set of axioms that resolve all cases of unsatisfiability. We demonstrated this across the OBO Foundry, and found that relatively few axioms can be removed to resolve all unsatisfiable classes.

While the algorithm removes a minimal set of axioms to make an ontology coherent, it does not repair the root cause of the contradiction. In one case we showed that a large number of unsatisfiable classes in the Gene Ontology were caused by a mistaken use of a parthood relationship. This cause for unsatisfiability was complex, but would have been revealed by an automated reasoner had the axioms of MIREOT-ed classes been included. This indicates that the unconstrained use of MIREOT has introduced a new challenge for ontology interoperability, which must now be addressed. The question remains, however, of how best to balance the challenges of developing ontologies with the hardware resources and tools available, while at the same time maintaining consistency and interoperability between ontologies. Our results illustrate how the unMIREOT tool can be used to help ontology developers identify problematic axioms in their ontologies, and explore them to diagnose causes of contradiction.

One approach to preventing contradictions is the integration of unMIREOT with the build process for biomedical ontologies. OBO uses a shared central build system which can be configured to validate ontologies against scripts that check for problems. By using a powerful build server to combine ontologies with the ontologies they refer to and check for inconsistencies before release, developers would be able to continue to use MIREOT while ensuring continuing compatibility. It is also possible that either the MIREOT or

OBO guidelines should be revised. Including the axioms of referenced classes would allow for local consistency checking with an automated reasoner. Because many axioms are inherited, and restrictions are placed transitively, the axioms of an entire ontology or at least a derived module would need to be imported. This could be recommended in the case of small, high-level ontologies such as BFO and RO, which should not cause performance or space issues. Without actually including the ontology in the imports closure, however, it would not solve the problem of sourcing ontologies becoming out of date with the ontologies they reference.

While we have shown that there are large clusters of unsatisfiability across the OBO Foundry, it is unclear whether or to what extent these issues are affecting ontology-based analysis techniques. Incorrect inferences could affect the results of gene enrichment analysis, inter-ontology phenotype mapping, semantic simiarity tasks, or any analysis that relies on ontology axiomatisation. We would like to explore this by implementing a reference task, and comparing the performance before and after repair. We would also like to investigate automatically determining sources of disagreement. While ontologies can be repaired by the unMIREOT program, and examination of its output can help to identify the root cause of unsatisfiability, this can still be a time consuming and complicated process. It's possible that algorithmic tools could be developed to aid ontology developers in identifying the actual cause of the inconsistency, or instead to create a set of minimally destructive axioms to remove from the ontologies.

CHAPTER 4

VOCABULARY EXPANSION

4.1 Introduction

We have previously discussed that there is a lot of research in the area of extending vocabularies for text mining, both with and without ontologies, and that extended vocabularies can improve the performance of text mining tasks. We also discussed that through concept re-use and sharing, particularly by means of the MIREOT system, ontologies can extend knowledge and metadata concerning a term originally defined elsewhere, in a linkable way: providing greater granularity to a class or description in a different context. In addition, we showed that some ontologies repeat definitions for entities in similar contexts, but with potentially different metadata.

No ontology-based approach to synonym expansion has attempted to use MIREOT-ed classes or alternative class definitions to obtain more synonyms. We gave the example of *hypertension* (HP:0000822), which is defined in the context of a phenotype, but is defined as a disease elsewhere (DOID:10763). There are also more specific and granular definitions of hypertension and related conditions available from the Hypertension Ontology (HTN)[69]. Particularly, HTN uses MIREOT to import and extend the HP class *hypertension* (HP:0000822), and adds extra information to it. For example, with the additional label 'hypertensive phenotype': a label not available in HP. In addition, while it contains the primary 'hypertension' label that HP defines, it does not include the additional syn-

onyms defined by HP itself. Furthermore, the subtle distinctions between concepts that biomedical ontologies capture do not necessarily matter for most text mining tasks, because they operate within a single context. In text-mining of electronic healthcare records, for example, the hypertension phenotype and the hypertension disease are functionally equivalent.

We hypothesise that because ontologies are constructed with different loci: different contexts, domain experts, and source material, ontologies that share lexical concepts and terms will contain different, but valid, synonyms for a particular context. By considering all of these concept descriptions, we can gain a greater set of synonyms that will improve the power of information retrieval and information extraction systems. In this chapter, we describe a synonym expansion approach that combines lexical matching and semantic equivalency to obtain new synonyms for biomedical concepts.

Lexical matching, mapping terms across ontologies via shared labels or metadata, has seen previous investigation, and is one of the major techniques used in ontology alignment[83]. It is this method by which strongly related classes with definitions in different contexts can be found. For example, moving from the HP definition of the hypertension phenotype, to the disease of hypertension in DO.

We have also shown that automated reasoners enable querying for class descriptions. Ontologies that extend the same concepts do not necessarily share the same label; this can happen due to the ontologies becoming out of sync, or due to limiting the copying of what are seen as unnecessary details for a particular purpose. Equivalency queries, however, can obtain these classes, as well as any other classes inferred as equivalent to the target class, and then extract and consolidate their metadata. In the hypertension example above, additional labels would be found from DO via lexical matching, while additional examples would be found from the MIREOT class defined in the HTN ontology.

We present an algorithm that uses these two approaches to expand synonyms for ontology terms. We implement them into a software tool, and evaluate it in several ways. First, we extend synonyms for all non-obsolete phenotypes described by HP. We

then manually validate the candidate synonyms for a random selection of the synonyms selected from HP. We also use the subset of cardiovascular phenotypes to evaluate the amount of information the expanded synonyms returns for a clinical text annotation task and an information retrieval task over biomedical literature.

4.2 Materials and Methods

OWL ontologies use a number of conventional annotation properties to define labels and synonyms. These span a range of confidence and degree of synonymy. In this chapter, we consider frequently used annotation properties, summarised in Table 4.1. These are the annotation properties consolidated into the 'synonym' property by the AberOWL API. Another oboInOwl synonym, *hasRelatedSynonym* excluded, because the labels provided by these synonyms are too loose.

We developed an algorithm using the AberOWL API to consider semantic and lexical matches for a given list of classes, and output a list of synonyms for each. It makes use of the name query and semantic query functions, documented at `http://www.aber-owl.net/docs/`.

Using this algorithm, we performed an expansion of synonyms for all non-obselete subclasses of *Phenotypic abnormality* (HP:0000118) in the Human Phenotype Ontology (HP). We use this set of classes because it contains terms relevant for phenotyping patients from text documents. The ontology was downloaded and expansion performed on the 21/04/2019.

To evaluate the performance of the algorithm, we randomly selected 500 classes from the expanded version of HP for manual validation. Synonyms already asserted by HP were removed from the set, because they were already assumed to be correct, and would not contribute to measuring the performance of the synonym expansion algorithm. A clinical expert (WB) marked each synonym as correct, incorrect, or ambiguous. The expert was asked to answer correctly or incorrectly on the basis: "if a patient has *synonym*, would

Table 4.1: Summary of conventionally used annotation properties considered in this experiment. Definitions come from the description of the annotation properties in their respective top-level ontologies.

Annotation Property	**Identifier**	**Definition**
label	rdfs:label	"a human-readable version of a resource's name[9]."
altLabel	skos:core#altLabel	"An alternative lexical label for a resource[11]."
has_exact_synonym	hasExactSynonym	"An alias in which the alias exhibits true synonymy[20]."
has_narrow_synonym	hasNarrowSynonym	"An alias in which the alias is narrower than the primary class name. Example: pyrimidine-dimer repair by photolyase is a narrow synonym of photoreactive repair[20]."
has_broad_synonym	hasBroadSynonym	"An alias in which the alias is broader than the primary class name. Example: cell division is a broad synonym of cytokinesis[20]."
alternative term	IAO_0000118	"An alternative name for a class or property which means the same thing as the preferred name (semantically equivalent)[14]."

it also be true that they have *original label*?" Entries were marked as ambiguous if the synonym was in a different language, or the validator otherwise did not have enough knowledge of the phenotype to determine whether or not the synonym was correct.

We also investigated whether the expanded set of synonyms improved the power of text mining tasks, by comparing the output of an information retrieval and an information extraction task before and after its input vocabulary was expanded. We used Stanford CoreNLP's RegexNER annotator to annotate 1,000 randomly sampled entries from the NOTEEVENTS table in MIMIC-III (MIMIC)[57]. MIMIC is a freely available healthcare database, containing a variety of structured and unstructured information concerning around 60,000 admissions to critical care services[80]. We annotated the sample with all subclasses of *Abnormality of the cardiovascular system* (HP:0011025), comparing the number of annotations before and after synonym expansion. This investigation was

performed on 17/01/2020.

Using the same set of subclasses of *Abnormality of the cardiovascular system* (HP:0011025), we compared the sum of article counts returned for a disjunctive query of all labels and synonyms for each term, before and after synonym expansion. MEDLINE is a searchable database of literature metadata in the life sciences, containing more than 25 million article references[7]. MEDLINE was queried on the 27/01/2020.

The synonym expansion tool is available standalone, online at `https://github.com/reality/expand_terms`, or as part of the Komenti semantic annotation tool, which is available at `http://github.com/reality/komenti` [143].

All files described in the validation files (excluding the MIMIC-III data files), along with the commands necessary to repeat the experiments are available at `https://github.com/reality/synonym_expansion_validation/` [129].

4.3 Results

The synonym expansion algorithm performs the following process, for each class provided as input (in this context, 'every ontology' is every ontology contained in AberOWL):

1. Extract the labels and synonyms of any classes in any ontology with a label or synonym that exactly matches a label or synonym of the input class.
2. Run an equivalency query against every ontology using the IRI of the input class, extracting labels and synonyms for any classes returned.
3. Of the candidate synonyms produced by the first two steps, discard any that were:
    - Defined in ontologies that were found to produce incorrect synonyms.
    - Have the form of a term identifier.
    - Contain the input class label as a substring.

Some ontologies include term identifiers as labels, and these are unhelpful for text-mining applications (at least in this context). Therefore, candidate synonyms that con-

tained a colon or underscore were removed. We also found that certain ontologies reliably produced incorrect synonyms: GO-PLUS[71], MONDO[138], CCONT[52], and phenX[62]. Several of these ontologies are meta-ontologies automatically constructed from several ontologies using alignment methods, and it's possible that errors in that process were the cause of the incorrect synonyms. Certain annotation properties were also incorrectly detailed by the AberOWL API as being labels, such as *europe pmc* and *kegg compound*; it is unknown whether this is a fault of AberOWL or of the source ontologies. Candidate synonyms defined by the problematic ontologies or matching the problematic annotation properties are automatically removed.

4.3.1 Human Phenotype Ontology Expansion

We applied the vocabulary expansion algorithm to all 14,406 non-obsolete subclasses of *Phenotypic abnormality* (HP:0000118) in HP. HP itself asserts 29,805 labels and synonyms. The number of labels and synonyms following expansion was 54,765. Therefore, the algorithm found 24,960 additional synonyms across HP.

For the term *hypertension* (HP:0000822), 40 unique synonyms were found by the algorithm, across 7 source ontologies. 2 of these synonyms were asserted by HP itself. The final number of synonyms was eventually reduced to 33, following the disclusion of terms that contain the original 'hypertension' as a substring. The synonym sources are shown in Table 4.2. These include several general phenotype ontologies, as well as domain specific ontologies such as the Hypertension Ontology (HTN), and the Cigarette Smoke Exposure Ontology (CSEO).

4.3.2 Evaluation

Table 4.3 summarises the results of the manual validation. The manual validation revealed that many synonyms returned were actually given in non-English languages. While OWL ontologies do allow for parameters that distinguish which language the property is in,

Table 4.2: Source of the 40 synonyms found for the term 'hypertension' per-ontology.

Ontology	Number of Synonyms
CCTO[93]	9
CSEO[175]	9
DO[136]	8
HP[87]	2
HTN[69]	13
WHOFRE[115]	9
NCIT[141]	8

AberOWL does not index them. Therefore, it is not currently possible to distinguish between English and non-English synonyms.

Through analysis of the false positives, we found that many were caused by errors in the ontologies that synonyms were sourced from. Several synonyms for *motor aphasia* (HP:0002427) were marked as incorrect. They were incorrect because they refer to dysphasia, for example "Broca Dysphasia." Aphasia and dysphasia are different conditions: the first refers to a partial loss of language, and the latter to a full loss of language. All of these incorrect synonyms were sourced from *Aphasia, Broca* (MESH:D001039) in MESH.

Table 4.3: Metrics for clinical expert validation of 866 generated synonyms for 500 terms. Synonyms already included in HP were not included in the validation. Synonyms were marked ambiguous if not English, or if the validator did not have enough expertise to confidently judge it.

Terms	Total Synonyms	TP	FP	Ambiguous	Precision
500	866	613	59	195	0.912

The two text mining evaluations used labels and synonyms for all non-obsolete subclasses of *Abnormality of the cardiovascular system* (HP:0011025). HP asserts 2,205 labels and synonyms for these classes, while the expanded set of labels numbers 5,336. The results are summarised in Table 4.4.

Table 4.4: Amount of labels for *Abnormality of the cardiovascular system* (HP:0011025) before and after synonym expansion, and results of the two text mining tasks using them as vocabularies. MEDLINE results are the sum of the number of results returned by each query.

Vocabulary	Labels	MIMIC-III Annotations	MEDLINE Results
HP Labels	2,205	1,104	8,191,564
Expanded HP Labels	5,336	1,447	13,513,342

4.4 Discussion

We have demonstrated that AberOWL's lexical search and semantic query functions can be used to enrich ontology vocabularies by interpolating synonyms and labels from the wider ontology ecosystem. In a hypertension example, we showed that very specific disease-level ontologies can contribute additional vocabulary. In that case, the HTN ontology contributed 13 new synonyms to the HP term. By automatically leveraging these, as well as synonyms from other generalised terminologies, we can effectively enrich vocabularies for biomedical terms.

A manual validation of a selection of expanded synonyms across HP showed a high precision. However, it revealed that AberOWL cannot distinguish between the languages of term labels, and therefore many non-English labels were returned as synonyms. Because ontologies define the language of labels, this feature can be added, and then the algorithm can be modified to permit only synonyms of the same language as the input. This could also be controlled partially bydiscounting additional ontologies from results. For example, WHOFRE is actually a non-ontology mapping of French vocabulary to UMLS. Analysis of false positives also revealed errors in external ontologies. Ultimately, this approach will inherit any such errors. This effect can be controlled by extending the list of discluded ontologies, although this might reduce the number of true positive synonyms found. For any uses where a reduced precision is not acceptable, candidate synonyms should be checked by a domain expert.

We also showed that vocabularies expanded by our method increase the amount of data returned by two information retrieval tasks. We have not, however, shown whether

the extra information returned is actually useful or relevant for a particular purpose. We can assume, from the manual validation, that some of the additional data returned are incorrect, though most should be correct.

The most important potential limitation of the work is that it violates the notion that the IRI of a concept uniquely identifies it, rather than its name. This is because OWL ontologies do not follow the unique name assumption. Theoretically, false positives could be generated by a lexical match on a homonym, which then has different synonyms itself. We believe, however, that this effect should be limited in the case of highly specific biomedical language. Furthermore, any error is mitigated in practice by dataset context limitation. For example, synonyms from another context incorrectly associated with a medical concept are unlikely to be found in clinical letters.

False synonyms could also be removed on the basis of a corpus search: for example, if a candidate synonym never or rarely appears in the same document as another label for that term across a literature corpus, it's possible that it refers to a different concept from a disjoint context. This could also be performed by analysing the metadata of text corpora: for example, if two terms are never or rarely associated with literature from the same journals, the same field, or the same content tags, it's possible they have different meanings. However, the success of such approaches could be limited, since they are at odds with the notion of consistent use of language within singular articles or contexts. We would also like to investigate whether this could be achieved using word embeddings. We would also like to explore database cross-references as a potential source for additional synonyms. These cross-references establish a semantic link with entries in non-ontology databases. In biomedicine, frequent mappings include DrugBank, UMLS, ICD, and SNOMED. These associations provide further opportunity for vocabulary expansion. This functionality could also be integrated into the AberOWL platform, by indexing non-ontology databases and associating these with the cross-references, or otherwise making them accessible via the API.

CHAPTER 5

NEGATION DETECTION

5.1 Introduction

We previously discussed that the use of context disambiguation tools, particularly negation detection, can improve the performance of information extraction tasks. Previous approaches to negation detection include regular expression or hotspot based methods, machine learning, and dependency resolution rules.

While dependency resolution methods have proven powerful, one of the limitations of existing approaches, despite their ability to make use of the rich grammatical model of a sentence, is that they use complicated grammatical rules requiring expertise and development time, and do not generalise well. The inherent complexity and ambiguity of human language leads to such a variety of grammatical models for sentences that no satisfactory set of rules can be determined via manual training over a small set of sentences.

We hypothesise that a more general rule-based approach to typed dependency negation detection will perform and generalise better than rule-based approaches. We propose an algorithm that uses typed dependencies, but avoids defining specific patterns of dependency. In this way, it should require minimal training, and provide consistent high performance across medical datasets. The algorithm proposed by Gkotsis et al. is the closest to the proposed algorithm, however it uses a graph pruning approach, removing

subordinate clauses and irrelevant intermediate nodes, before employing what is essentially a string search for negation vocabulary[55]. It reports itself as performing similarly to ConText, with a slightly higher recall. In an independent investigation, Manimaran and Velmurugan found that it performs extremely well, outperforming other popular methods, when extended with a richer negation vocabulary[101].

In the proposed algorithm, we instead measure 'dependency distance:' the distance in a typed dependency graph between a negated term and the target term, as the measure of negation context. In this way, we mirror the generic and transferable 'hotspot' method employed by NegEx and ConText, while extending it with the notion of grammatical relatedness afforded by dependency modelling.

In this chapter, we present the negation algorithm, and evaluate its performance on two medical corpora, in comparison with a number of different negation detection algorithms. First, against the MIMIC-III critical care database[80], and then against clinical letters mentioning HCM at University Hospitals Birmingham (UHB).

5.2 Materials and Methods

The negation algorithm is implemented in Groovy, and makes use of the Stanford CoreNLP dependency resolver. It is available as part of the Komenti text-mining tool, at `http://github.com/reality/komenti`. Evaluation text was annotated with Stanford CoreNLP's RegexNER, also using the Komenti tool. Annotation used all non-obsolete subclasses of *Phenotypic abnormality* (HP:0000118) in HP as vocabulary.

5.2.1 Corpus Generation and Training

The MIMIC dataset was derived from the MIMIC-III critical care database. The text was sampled from the NOTEEVENTS table. Entries were sampled randomly, and then split into sentences. A random sentence was then selected. 500 randomly selected sentences were used for training. Training involved annotating the 500 sentences with biomedi-

cal concepts, running the negation algorithm against the set, and examining error cases to identify additional negatory vocabulary not currently included in the software, and identify any errors in the evaluation software, or counting algorithms. To ensure fairness, extra vocabulary terms were also added to the other algorithms evaluated in this experiment (if not already present). In the case of NegBio, only grammatical rules are accepted. Therefore, for each of the two negatory words that were missing: deny and not (surprisingly), a bi-directional rule was introduced. It's possible that more finely tuned rules could have been produced for better performance, but the only training considered in this experiment is the addition of extra negatory words. The NOTEEVENTS table was sampled again to obtain 7000 sentences for a test set. These were annoatated with HPO terms using the Komenti tool, yielding 1,300 annotations. HPO query and sampling were both performed on 28/12/2019.

In the case of the hospital validation (HCM), 5000 sentences were sampled from a clinical text corpus of documents matching HCM keywords. The construction of this corpus is explained in more detail in the next chapter. To sample the corpus, a file was selected at random, and then one sentence randomly from that file, repeating the process until there were 5,000 sentences. The sentences were annotated with HPO terms using the Komenti tool, yielding 1,077 annotations. No training set was used for this dataset, to test algorithm generalisability.

In both cases, during sentence selection, selection criteria were used to constrain the text returned. This was for two purposes. First, to ensure that narrative text was returned, rather than field-based, table-based, or irrelevant text. Second, to limit the length of sentences to make it easier to perform manual validation. Sentences shorter than 4 words and longer than 30 words were excluded, and sentences containing phrases indicating field data were excluded. Sentences with indicators of nonsense (e.g. due to scanned documents) were also removed.

These problems could be solved by additional pre-processing of the text, but this task is not the subject of this investigation, and using shorter sentences should not advantage

any particular algorithm (although the dependency parsing algorithms are more sensitive to correct grammar). These parameters and pre-processing options were manually tuned, and decided upon during the training phase. For simplicity, where a single concept was mentioned multiple times in a single sentence, only one annotation was preserved, and negated concepts were given priority. This is potentially a small source of error, but should not favour any particular algorithm. The test code was designed to ask, in each case, "is an instance of the word negated in this sentence."

In both cases, all annotations were manually labelled with respect to their negation status, determining whether the annotated concept was negated in the sentence. In the case of ambiguity, the concept was marked as negated if the patient doesn't have the condition, and not negated if they do have the condition. This is because the purpose of the negation detection algorithm, in this context, is the exclusion of concept mentions from evidence of a patient having a condition if they do not have it. The negation labels were checked by a clinical expert (WB).

5.2.2 Evaluation

In choosing negation algorithms to compare with, we found that many are not public software. For reasons discussed in the introduction, we did not consider machine-learning methods. The algorithms tested are NegEx, pyConTextNLP, negation-detection, NegBio, and Komenti (the proposed algorithm). More information about the algorithms, including version numbers and sources, are included in Table 9.1 (supplementary).

While these are all, in some sense, rule-based classifiers, we make a distinction between trigger based classifiers, and dependency-based classifiers. Trigger based classifiers define a set of regular expression rules, that define a negatory construct. For example, PyConTextNLP includes the following rule:

```
Comments: ''
Direction: forward
```

```
Lex: without sign
Regex: without sign(s)?
Type: DEFINITE_NEGATED_EXISTENCE
```

In the case that a *Regex* property is defined, the regular expression is used to match a rule: in this case a grammatical form that expresses the text 'ruling out' a subject with respect to a concept. In the case that this property is not defined, a definite match is made using the *Lex* property. In both cases, the *Direction* property stipulates whether the negatory construct (pattern or absolute phrase) should appear before or after the concept. The algorithms also determine their own heuristic for how far the concept should be from the negatory construct in the sentence.

Dependency-based algorithms instead use a small dictionary of negatory words, in combination with a grammatical parser that produces a dependency graph model for a sentence. An algorithm is applied to that graph to determine whether a negatory construct applies to a particular concept. If an algorithm raised an error during processing of a sentence, the result was taken to be false (i.e. the concept was not negated). This happens in the case of NegBio, for example, when a parse tree cannot be constructed for an input sentence.

We also sought to make a gold standard dataset with which to make future algorithm comparisons. Therefore, we examined the errors of the three best-performing algorithms by f-measure, and updated the manually annotated labels if incorrect. The results presented in this paper are those from the revised corpus.

The Linux command *time* was used to measure the execution time for each algorithm, with the 'real' measurement taken. Within each dataset, every algorithm was evaluated using the same machine, but two different machines were used for each dataset. There is an exception in the case of NegBio for the MIMIC-III validation, which would not run on the same hardware setup as the others, and therefore was evaluated on the same machine as the HCM dataset.

5.3 Results and Discussion

5.3.1 Algorithm

Algorithm 1 describes the algorithm that determines the negation of a concept in a sentence. The dependency resolution algorithm produces a typed dependency graph, which is passed to this algorithm as input. This graph is formed of nodes that represent word tokens, and edges that represent their grammatical relationships. Together, they form a grammatical model of the sentence. Each edge is labeled with a particular kind of relationship, such as negation or adjectival noun modification. The edges also have a direction, that define a **governer** and **dependent** for each relationship. For example, in a noun modification relationship between the words **light** and **touch**, the governer would be touch, and the dependent light, because the subject is the noun 'touch', while light is its modifier. The graph can aso be thought of as a set of assertion triples: a governer (subject), dependent (object), and relationship (predicate).

The basis of the algorithm is an attempt to find a transitive relationship between a negation construct and a word of interest. Because the typed-dependency graph does not support multi-word nodes, sentences are pre-processed to replace the concept of interest with a single neutral word, such as 'biscuit.' The algorithm then identifies whether either a negatory word or a word with a negatory dependent exists in the path to the root grammatical relationship of the graph. A negation vocabulary is also used in addition to negatory relationships, because the dependency resolution algorithm does not reliably identify all negatory constructs with a negation dependency (for example, with the word 'exclude'). Other sub-graphs are not explored, because these separate paths contain negatory constructs that refer to other objects in the sentence, and it is therefore a useful splitting factor. If a match is found, its distance from the target concept is then measured. This relationship distance heuristic is used to eliminate unrelated negatory constructs that refer to other words. The cut-off point for this parameter can be modified, but is set to 4. This value was manually chosen during algorithm development process.

```
Data: S = A typed dependency graph of a sentence.
T = Tokenised form of the concept of interest
V = Vocabulary of negation words.
Result: True if the concept is negated in the sentence, False otherwise
x = node(T)
edges = getEdges(x)
negated = False
for e in edges do
    if predicate(e) == neg then
        negated = True
    end
end
if negated then
    return True
end
path = pathToRoot(x)
rDistance = 0
for node in path do
    rDistance++
    dependents = getDependents(node)
    negated = False
    if word(node) in V then
        negated = True
    end
    for rel in dependents do
        if word(rel) in V then
            negated = True
        end
        if predicate(rel) == neg then
            negated = True
        end
    end
    if negated and rDistance < 4 then
        return True
    end
end
return False
```

Algorithm 1: Algorithm for determining the negation of a concept in a sentence.

The algorithm also has a preprocessing step. If the sentence contains one of the words in the negation vocabulary, followed by whitespace, followed by the concept of interest, this is transformed into its own sentence with the word 'excludes' appearing directly before the concept. This is because there is a tendency for the CoreNLP dependency resolution algorithm to express such grammatical forms using negatory words other than 'excludes' as adjectival, rather than negatory, relationships.

5.3.2 Evaluation

We discovered during the training period that both the negation-detection and NegBio algorithms were not properly able to handle parenthesised text, causing a lot of error. This was easily fixed by transforming any parenthesised text into a new sentence (i.e. by replacing each bracket with a period and a space). The sentence containing the concept of interest is then chosen for negation analysis. This modification was implemented, and the results from the modified algorithm are provided under the algorithm names *negation-detection (parfix)* and *NegBio (parfix)*.

Table 5.1 summarises the result metrics for both the HCM and MIMIC datasets. In both cases, the best performing algorithm was Komenti, with respect to its f-measure. However, it was outperformed in both precision and recall by other algorithms in both corpora. In the case of MIMIC, the *negation-detection (parfix)* algorithm comes very close by f-measure, though it suffers lower precision and recall than Komenti. For HCM, the NegEx algorithm comes extremely close via f-measure, but has a slightly lower recall.

With respect to generalisability, the Komenti algorithm also has the smallest magnitude of f-measure difference between the two corpora, of 0.019, closely followed by negation-detection (parfix) at 0.024, and NegBio at 0.045. The NegEx and pyConTextNLP algorithms performed much better on the HCM corpora, while NegBio performed much worse.

Another important factor is running time. The quickest algorithm in both cases was NegEx, finishing in less than 2 seconds. pyConText is also very quick, finishing in less

Table 5.1: Performance comparison of negation algorithms on sentences sampled from MIMIC and HCM datasets. The best performance for each metric in each dataset is emphasised.

Corpus	Algorithm	Precision	Recall	F-measure	Time
MIMIC					
	NegEx	0.674	**0.948**	0.788	**0m1.699s**
	pyConTextNLP	0.467	**0.948**	0.626	0m55.739s
	negation-detection	0.584	0.657	0.619	53m17.757s
	negation-detection (parfix)	0.834	0.91	0.87	21m48.981s
	NegBio	0.82	0.471	0.598	29m11.643s
	NegBio (parfix)	**0.88**	0.665	0.757	42m57.643
	Komenti	0.844	0.942	**0.89**	1m0.149s
HCM					
	NegEx	**0.905**	0.905	0.905	**0m1.082s**
	pyConTextNLP	0.85	**0.931**	0.889	0m41.446s
	negation-detection	0.898	0.889	0.894	4m19.412s
	negation-detection (parfix)	0.898	0.889	0.894	4m10.124s
	NegBio	0.678	0.611	0.643	39m4.309s
	NegBio (parfix)	0.711	0.611	0.657	44m18.804s
	Komenti	0.893	0.926	**0.909**	0m38.225s

than a minute in both cases. Komenti finished in just over a minute in the slower case. While the HCM dataset was slightly smaller, its evaluation was performed on a much more powerful machine. This is reflected in the difference between the running times for negation-detection on the two sets. NegBio, however, remains slow, taking around 40 minutes in both cases. Upon investigation, we found that negation-detection makes use of multiprocessing, while NegBio does not. It is also curious that in the case of MIMIC, the parfix modification more than halved the running time of negation-detection, but actually increased the running time of NegBio. This is surprising, because smaller sentences should be quicker to parse, and perhaps suggests that the running time is linear with respect to the number of sentences, rather than being dependent on complexity of grammar.

While most algorithms maintained their performance across corpora, Negex and pyConTextNLP (which use the same basic algorithm) showed much better performance against HCM, and NegBio much poorer performance. In fact, NegBio maintained its relatively poor recall, but lost its precision. The MIMIC dataset used a lot of sentences in the form 'not x,' some of which were parenthesised (accounting for the increase in

performance from the parfix variant), as well as sentences describing patients denying symptoms. As described in the methods, we had to add a bi-directional rules to capture the 'not x' and 'denies x' forms to NegBio during the training for MIMIC, because it was not caught by default. We can surmise from its lack of generalisation to the HCM dataset, that it contained negation forms that were neither caught by these additions, nor its in-built grammatical rules. Overall, however, the language in the HCM dataset was simpler. There were fewer run-on sentences containing many observations without punctuation separating them, and we expect that this is the reason the NegEx algorithms showed better performance over this corpora.

Our results show that in addition to trigger based algorithms, some dependency resolution algorithms, specifically those which use general heuristics rather than grammatical patterns, also generalise across datasets well. In addition, they perform well in situations that NegEx algorithms do not. Nevertheless, a training phase is still necessary. Training over MIMIC identified three additional negatory words to add to the vocabulary. This process is relatively easy and quick when compared to the development of grammar rules, however.

Some algorithm errors could also be mitigated by pre-processing. For example, by transforming sentences that appear like fields or tables. This would, however, also be part of an involved training process, and any changes would not necessarily generalise well. This is evidenced by the relatively small difference in performance for parfix variants over the HCM dataset compared with the dramatic improvements over MIMIC. Another potential source of dataset bias is that certain phenotypes are over-expressed in medical texts, and therefore the samples used in this experiment will be testing the negatory language used around them much more than other phenotypes. A clear example of this is pain, which accounts for 127 of the 1300 annotations in the MIMIC dataset. Meanwhile, the HCM dataset were documents discussing HCM, which carries with it a standard set of phenotypes and comorbidities.

The main source of error that remains for Komenti and negation-detection is the

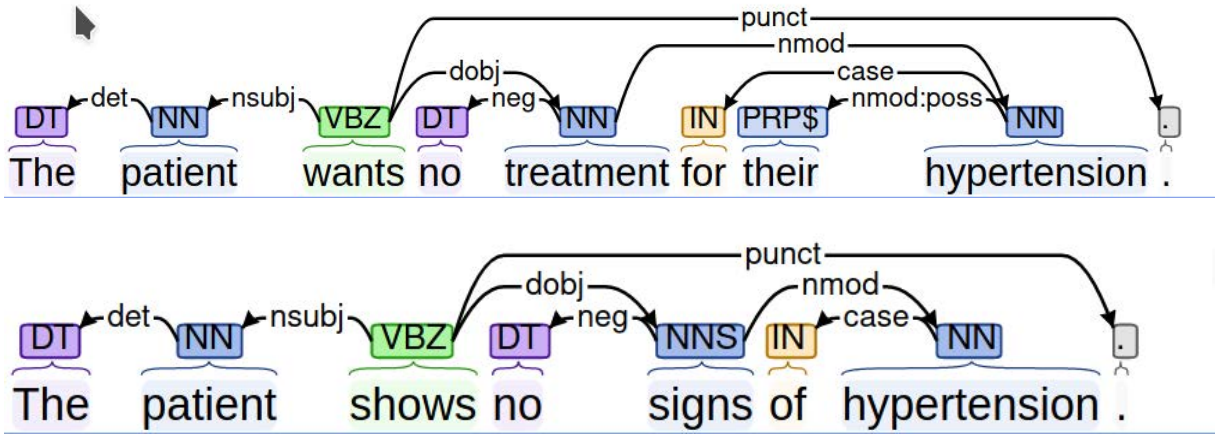

Figure 5.1: Two sentences concerning hypertension, with typed dependency annotations.

problem of whether the negation of intermediary phrases also negates the target noun. For example, in the sentences about hypertension shown in Figure 5.1, the negatory modifiers actually apply to intermediary nouns (and these are the direct objects of their sentences). These nouns are 'treatment' and 'signs' respectively, and they are connected to hypertension by a noun modification relationship. The grammatical relationship between the negator and the concept of interest is the same, but they express very different things. The first talks about treatment of hypertension, the negation of which does not indicate that the patient does not have hypertension (rather the opposite), while the second talks about signs of hypertension, which if negated also indicates there is no hypertension. This difference can also depend on the verb used, and many other potential expressible constructs.

This problem is not easily solved. The Komenti and negation-detection algorithms are completely ignorant to this kind of relationship, while NegBio only understands codes for the predicates themselves (so it could not tell the difference between a 'sign' and a 'treatment'). These algorithms could be modified to accept patterns of different intermediary nouns, indicating whether or not its negation also negates the target concept, and indeed these could be implemented using NegEx regular expressions. Alternatively, a simple classifier could be trained using annotated texts to learn whether or not these negations apply transitively. However, both of these approaches would introduce a complexity and

necessity for training that betrays the purpose of a simple heuristic-based algorithm.

Nevertheless, this in combination with ambiguous and badly structured text (such as tables and forms) are the primary source of error for the algorithms in this experiment. In the next chapter, we will suggest a method for overcoming this source of error in a practical system.

In the future, we would like to develop a method of automatically tuning the node distance parameter. Its optimal setting potentially depends upon several features of the free text the algorithm is employed upon, including the complexity and domain of the language expressed. Moreover, heuristics within the individual sentences could be used to tune the parameter: the length of the sentence, the total depth of the target concept within the grammatical model, and the total number of noun class words in the sentence. However, the aim of this work was to develop a base method for negation using co-reference models; further development of heuristics that depend upon specific sentence structures risk suffering a large number of edge cases and an inability to easily generalise.

Another limitation of the algorithm is that it currently does not support multi-word negatory constructs. Figure 5.2 shows an example of such a rule in pyConTextNLP. The Komenti cannot currently model this kind of relationship, as the grammatical tagger does not recognise 'can rule out' as a negatory phrase (although it can recognise some multi-word entities, or represent them as noun phrases), and the dictionary is matched against each token (word) individually. While this has not proved to be a problem for the investigation described, it would be a desirable feature for further improvement of performance, and application to datasets where negation is expressed in more complex ways. It should be possible using the *entitymentions* CoreNLP plugin, which allows for the parsing of multi-word tokens.

In the introduction, an advantage of rule-based algorithms was given that they are explainable. Furthermore, a distinction was made between trigger and dependency based rule-based algorithms. In both cases, one can determine the reason a decision was made. In the case of trigger classifiers, such as NegEx, this is a matter of finding the trigger

```
Comments: ''
Direction: forward
Lex: can rule out
Regex: ''
Type: DEFINITE_NEGATED_EXISTENCE
```

Figure 5.2: Example of a pyConTextNLP rule.

rule that applies and examining its application. In the case of heuristic classifiers, this is somewhat simplified by there only being a heuristic rule, and a small number of negation modifiers. However, a disadvantage in the case of both of these algorithms is that the data model they operate upon, the grammatical reference graph, is not readily available after the fact. The algorithm could be modified to output a sub-graph or graph annotation that shows the reason with the decision was made. Such models could be presented in validation applications, and even be used as part of tools enabling the reactive development of corrective changes or rules.

We have shown that that the algorithm maintains performance over two datasets, across two different dialects of English, and two different healthcare settings: critical care and clinical noting. We have not, however, tested more than these environments, and we have not tested whether the algorithm maintains performance across different kinds of biomedical text such as literature, which may describe different or more complicated kinds of negation.

CHAPTER 6

PATIENT IDENTIFICATION AND PHENOTYPE EXTRACTION

6.1 Introduction

In this chapter we consider the work presented in the previous two chapters as components of a larger information extraction pipeline. We apply this pipeline to finding and phenotyping HCM patients across a clinical text record. We hypothesise that by extracting this information from the text record, we can come to a greater and more granular understanding of the HCM cohort than is contained in the structured databases at the hospital alone.

The synonym expansion algorithm is used in combination with expert advice to create a vocabulary for named entity recognition. These phenotypes are then annotated in text. We also create a variant of the negation algorithm that uses an uncertainty vocabulary, allowing for detection of ambiguity or uncertainty. Using this in combination with the previously described negation algorithm and presence of family-related work, we tag mentions of phenotypes with additional contextual information. We hypothesise that the use of several sentences and their associated classifications as input to document-level classifications for a concept will overcome the precision limitations of sentence-level classifications discussed in the previous chapter. Because concepts are mentioned several times in text, and most are classified correctly, evidence for the correct outcome should

outweigh small numbers of incorrect sentence-level classifications.

We then develop a classifier that uses all of these mentions and their associated relational context to classify patients with respect to the concepts. In this way, the pipeline forms an information extraction tool that phenotypes patients across an entire clinical document record. We evaluate the performance of the pipeline in several ways, both manual and automatic.

6.2 Materials and Methods

This work was undertaken at the Queen Elizabeth Hospital site of University Hospitals Birmingham NHS Foundation Trust (UHB) in the West Midlands, UK. There is a long standing service for HCM patients at the site, which includes a specialist clinic and rare disease registry that has been running for four years, since 2016. The documents were not de-identified as this was a service improvement project, where the clinical expert involved intends to follow up those individuals who have been lost to discharge. Only information relating to the concepts of interest for each study was extracted, with the remainder discarded. The hospital identification number was used to link documents belonging to the same patient and associated data to the registry, and other databases for validation.

6.2.1 Concept Vocabulary

A list of words and phrases were developed by a cardiology and HCM specialist (WB) that indicate hypertrophic cardiomyopathy, atrial fibrillation, or heart failure in text. These terms were then linked to HPO terms, and the synonym expansion algorithm was used to further extend the vocabulary.

6.2.2 Corpus Generation

Documents were obtained from the UHB patient document management system (OpenText, formerly Documentum) on 02/06/2019. The entire clinical document record, around 22TB in size, is contained in a series of PDF files. The corpus includes primary and secondary care referrals, clinic letters, discharge summaries, and digital noting. PDFs with words matching HCM keywords in the vocabulary were obtained. Only the most recent document mentioning each HCM, AF, and HF was used for analysis of each concept.

6.2.3 Annotation and Classification Pipeline

The annotation and classification pipeline was developed using Groovy, using Stanford CoreNLP[102]. The software used for the pipeline was a prototypal version of the Komenti software, with different annotation logic than can be found in the current tool. The current tool also does not yet include the exclusion matcher or overall status classification, although these features will be added in the future. Komenti is available at `http://github.com/reality/komenti`.

6.2.4 Training and Validation

A tool to validate results of the annotation was also developed, bundling a simple web server and client developed in NodeJS into a single-page web application. A training set of 300 patients was used over several iterations of the pipeline, using feedback from clinical validation via the tool to improve vocabularies and exclusion criteria. The negation vocabulary used was the one developed during the last chapter. The uncertainty vocabulary was initially populated with an dictionary developed for topic modelling[63], and was then curated during the training process. The exclusion vocabulary was developed entirely within this training process.

In the case of the manual validation, metrics are measured with respect to the pipeline's overall ability to assign the nature of the patient's relationship with the concept of interest:

uncertainty, negation, family history, positive. Family history is the only classification that is not mutually exclusive with the others, since it says something about the subject of the sentence, rather than the existence of the concept with respect to the subject. Manual validation results determined whether the algorithm had correctly assigned the status (uncertain, negative, family history, positive) of the patient with respect to each of the patients using the text provided. These results were used to inform modifications to the vocabulary and classifier, and validation was performed iteratively until a high precision and recall had been obtained on the training set.

As there is no single ground truth for diagnosis of phenotypes or conditions at the hospital, and because we are seeking to find new patients unknown to the structured data at the hospital, several methods of validation were used. These are summarised in Table 6.1. Since the cut-off dates for automated ECG reports and ICD-10 codes occur before the extraction date, patients for whom the extraction considered documents dated after these respective cut-off dates were excluded from the validation.

Table 6.1: Summary of the methods and resources used for final validation of the information extraction experiments.

Source	**Description**	**Collection Date**
Clinico-genomic Registry	Used to collect data on all HCM patients at the point of care in the specialist clinic since 2015.	02/06/2019
Manual	Validation performed via our tool on a further 300 patients not included in the training set.	N/A
Automated Electrocardiogram (ECG) Reports	Atrial fibrillation results were evaluated against a database of automated ECG reports produced by the machines. The reports contain a simple list of inferred conditions, and simple presence of the string 'atrial fibrillation' was used to mark patients as positive for AF.	26/06/2018
Hospital Episode Statistics (HES)	A database of ICD-10 codes at the hospital manually curated from letters describing inpatient stays.	31/03/2019

Many clinical letters, particularly those in the Accident & Emergency department,

exist only as scans of handwritten notes, marked forms, or screenshots of spreadsheets. Reading these documents is not in the scope of this work, but the data they contain may be reflected in ICD-10 codes or the rare disease registry, because they are formulated by human curators. Otherwise, ECG machine read-outs and other test results may contain phenotypes not discussed in the text record. Furthermore, we only analysed documents which mentioned HCM, which may not represent the full set of documents that discuss HF and AF. To ensure that we are measuring the ability of the pipeline to extract and correctly classify the phenotypes expressed in the text, the validation only counts patients for whom at least one mention of the concept being measured was found.

6.3 Results

6.3.1 Pipeline

Figure 6.1 describes the text extraction pipeline. Vocabulary construction takes place after determining a target cohort, by working with a clinical specialist to determine words or phrases that are indicative of the concepts of interest in the target corpus. This involves a distinction between dependent terms and additional phenotypes: the first being a term or terms that are necessary for a patient to be included in the dataset, and the latter being additional phenotypes of interest for patients whose record mentions the dependent concept. Vocabulary is also defined for annotation classification, defining words or phrases that would make a sentence irrelevant, refer to a family member, negated, or uncertain. The way these vocabularies are used is described in more detail in the next section.

After terms are defined, their labels are expanded by the algorithm defined by the *Vocabulary Expansion* chapter. These are not validated at this stage. In the case of the HCM investigation, the dependent concept is hypertrophic cardiomyopathy.

Following vocabulary construction, the clinical text record is searched for documents that mention dependent terms, using simple string matching. These documents are then

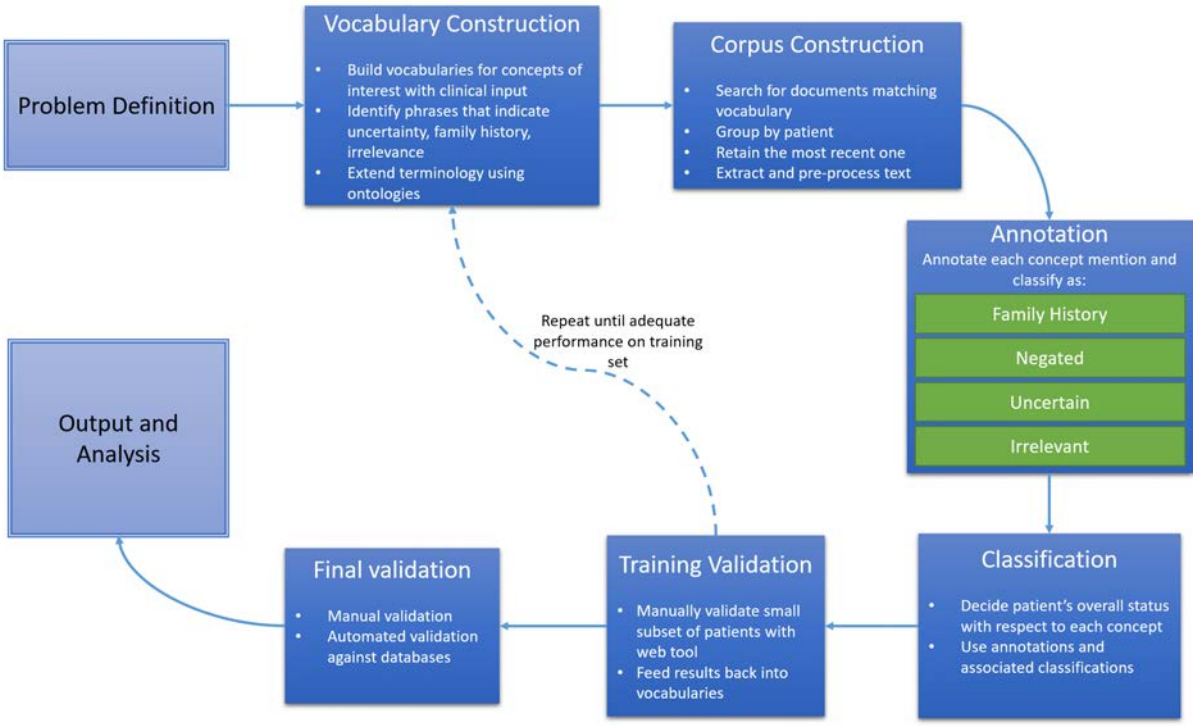

Figure 6.1: Flowchart and description of the pipeline to identify and phenotype patients from a clinical text record.

processed, extracting hospital numbers which are then used to group documents belonging to each patient. Then, the most recent document mentioning each concept of interest is kept. The text is then extracted from the documents, and pre-processed with basic NLP techniques, regular expressions, and other rule-based string manipulations. This extracted and pre-processed text forms the corpus.

During annotation, the text is split into sentences, and annotated with the vocabulary describing concepts of interest. In every case, the entire sentence is classified with respect to whether the concept in that sentence is negated, uncertain, or irrelevant. If it's considered irrelevant, it is thrown away. Sentences are also classified with respect to whether they refer to a family member or a family history, rather than the patient themselves. These sentence-level classifications are described in the next section, but they are used to decide the patient's overall status with respect to each phenotype. Only patients who are found to have, or found to be uncertain with respect to the dependent concept, are classified for other phenotypes.

Following these initial classifications, a training phase is entered. A validation tool (which is described later) is used by a researcher and a clinical expert together to measure the performance of the algorithm on a small subet of patients classified by the system. Feedback from this algorithm is used to inform modifications to the pipeline. Most often, this involves modifying the various vocabularies used as input for classifiers. It can, however, also inform modifications to the pre-processing steps. The entire process is then repeated with the changes, and continues to be repeated, until a suitable level of performance has been reached. Then, the final extraction is completed, and a final validation is performed both manually and against any databases. The manual validation does not include patients who were validated during the training phase, while the automated validations do, because it is our intent to measure the success of the overall process (including training) at phenotyping the population.

6.3.2 Annotation and Classification

Documents are split into sentences and annotated. Then, each sentence mentioning a concept of interest is tagged using four sentence-level classifiers:

1. Irrelevance
2. Family History
3. Negation
4. Uncertainty

The negation algorithm is the one described by the chapter *Negation Detection*. The uncertainty detection algorithm is a variant of the negation algorithm that instead uses a vocabulary that detects uncertainty. The exact definition of uncertainty depends on the problem definition, but in the case presented it is used to determine cases in which you would not want to assume, from the content of the sentence, that the patient has the condition. Ultimately, its purpose is disambugative: as a 'softer' rejection than negation. The use of an extra classifier, instead of simply extending the vocabulary of the negation

detection algorithm, is because patients who are classified as uncertain may be viable for additional follow-up. For example, hypertrophic cardiomyopathy is a frequently misdiagnosed disease, and patients for whom uncertainty is expressed about their condition may warrant further investigation by a specialist.

The irrelevance and family history classifiers are based on simple string matching of phrases in a vocabulary. In the case of the family history classifier, the aim is to identify sentences in which a patient's family or family history are mentioned (and therefore that concepts in that sentence may refer to the family member, rather than the patient). The reason that these are done using simple string matching, instead of using a more complicated algorithm, is that they do not require the same level of discernment as uncertainty or negation, and so that they can support multi-word phrases. The irrelevance classifier is used as a pre-processing step to determine whether the sentence is worth classifying. It is mostly used to remove templated text such as "assess dvt risk" or nonsense from scanned text. It is also, however, used as a catch-all in case of negation or uncertainty constructs that span multiple words. All of these are considered to be 'unclassifiable' or irrelevant text.

In the case of irrelevance, the sentence is discarded and not classified by other sentence-level classifiers. For family history, the sentence is still classified with respect to uncertainty, negation, and irrelevance, because this can provide useful information. After all sentences are classified, the patient's status with respect to each concept is determined. The patient's family history status with respect to each concept is also decided, using only the sentences referring to family history. Algorithm 2 describes the decision process. If the concept was not mentioned at all, the result is null, and no predicate exists in the results about the patient's status with respect to the condition. This is because not mentioning a condition is semantically different to the text actually stating that the patient does not have the condition. The sum of negated and uncertain sentences may exceed the total sentence count because sentences may be simultaneously uncertain and negated.

Data:
total_mention_count = Total number of sentences mentioning the concept in the document.
uncertain_sentences_count = Total number of sentences mentioning the concept classified as uncertain.
negated_sentences_count = Total number of sentences mentioning the concept classified as negated.
Result: Affirmed if patient has the phenotype. Negated if it is ruled out. Uncertain if it's uncertain. No if all mentions are irrelevant. NULL if it's not mentioned.

```
if total_mention_count == irrelevant_mention_count then
 | return "No";
end
total_mention_count -= irrelevant_mention_count;
if total_mention_count > 0 then
 | if uncertain_sentences_count > 0 then
 | | if negated_sentences_count + uncertain_sentences_count >=
 | |  total_mention_count then
 | | | return "Uncertain";
 | | end
 | end
 | if negated_sentences_count > 0 then
 | | if negated_sentences_count + uncertain_sentences_count >=
 | |  total_mention_count then
 | | | return "Negated";
 | | end
 | end
 | return "Affirmed";
end
return NULL;
```

Algorithm 2: Algorithm for determining whether a patient has a particular phenotype on a document level. No is functionally equivalent to NULL, but is separated for the purposes of evaluation. All counts do not include sentences that are classified as family history.

Validation Tool

To allow clinicians and researchers to manually evaluate results, a web-based validation tool was created. An example screenshot is shown in Figure 6.2. Because the correctness of the different patient-level classifications are validated, it's possible to measure from these validations the performance both of the overall algorithm at finding patients, but also the performance of the classifiers for uncertainty and negation. Patient-level comments can also be provided by the validator to explain why an error was made, or to suggest an

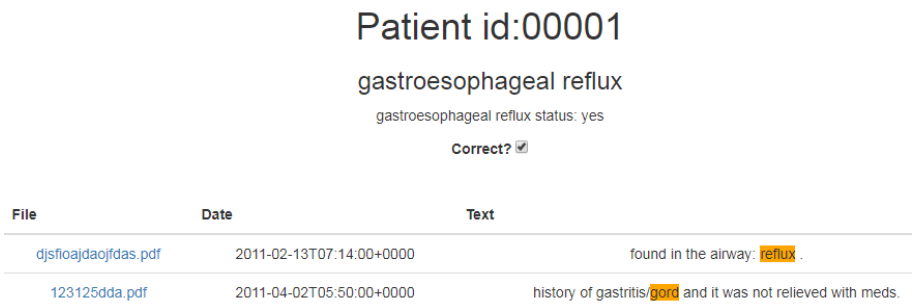

Figure 6.2: A screenshot of the validator tool used by researchers and clinicians to evaluate the performance of the algorithm, both for training and final validation. Sentence contents and ID numbers are synthesised. As evidence, sentences used by the algorithm to derive the overall diagnosis for the phenotype are listed. The filename is also given, and this forms a hyperlink to the PDF file itself, allowing the validator to consider the overall context in which the classification was made, which they can use to discern the ground truth.

upgrade to vocabularies.

6.3.3 Evaluation

Table 6.2 summarises the patients found and classified by the algorithm. 947 of the patients found were unknown to the rare disease registry. Table 6.3 shows that the prevalence of the comorbidities, HF and AF, was similar between the entire extracted set of patients, and the registry alone.

Table 6.2: Break-down of the classifications made by the algorithm, including the number of patients found by the extraction pipeline that were not already known to the rare disease registry.

Patients with HCM Keywords	3,120
Patients with only irrelevant HCM mentions	194
Individuals with uncertain HCM status	454
Individuals without HCM (explicitly excluded)	879
Individuals with HCM	1,787
Patients with HCM, unknown to registry	947

Table 6.4 show the results of the extraction evaluation. In all automated validations (registry, HES, and ECG) a high recall is shown, though it is higher for HCM than for the comorbidities. The results of the manual validation are further broken down in Table 6.5, showing precision values for each condition, and for each classification endpoint. For the

Table 6.3: The total number of patients determined by the extraction pipeline to have AF and HF, and a comparison of each complication prevalence amongst the full set of extracted patients, with its prevalence amongst the rare disease registry.

Complication	Extraction Patients	Extraction Prevalence	Registry Prevalence
Atrial fibrillation	288	20%	21.65%
Heart failure	161	11.18%	13.59%

overall counts, patient and family classifications are combined. Several classifications, such as family affirmation and uncertainty for HF and AF, had no occurrences, and so no metrics could be calculated. Items with very small numbers of samples carry extreme precision of 0 or 1, while in some cases precision was perfect even with a relatively large number of samples (e.g. family negation for HCM).

Table 6.4: Evaluation metrics for the pipeline. Precision and recall for validation of the pipeline results against multiple sources for each AF, HF, HCM, and overall. HES figures for AF and HF are only given for patients who were were affirmed or uncertain with respect to HCM, because only these patients were phenotyped. HF is not measured against the registry, due to the registry only recording whether a patient has had a HF admission. Precision is omitted for HES and ECG, as these are not a gold-standard resources, meaning the values would be misleading. Likewise, recall is not given for the manual validation, as it did not measure how many of true cases were found (since the validation only considered accuracy of classifications made on patients the pipeline found from the text), and the value would therefore be misleading.

Condition	Validation Method	Precision	Recall
HCM			
	Registry	0.47	0.901
	Manual	0.819	—
	HES	—	0.926
Atrial fibrillation			
	ECG	—	0.993
	Registry	0.793	0.87
	Manual	0.917	—
	HES	—	0.777
Heart failure			
	Manual	1	—
	HES	—	0.556
Overall			
	Manual	0.854	—
	HES	—	0.796

Table 6.5: Results of the manual validation for the HCM patient extraction. Overall counts include family history sentence assertions.

Condition	Classification	TP	FP	Precision
HCM				
	Affirmed	104	23	0.819
	Uncertain	24	5	0.792
	Negated	33	4	0.879
	Family (affirmed)	12	1	0.917
	Family (uncertain)	5	0	1
	Family (negated)	26	0	1
Atrial fibrillation				
	Affirmed	24	2	0.917
	Uncertain	2	0	1
	Negated	0	1	0
	Family (affirmed)	0	0	—
	Family (uncertain)	0	0	—
	Family (negated)	2	0	1
Heart failure				
	Affirmed	12	0	1
	Uncertain	3	2	0.6
	Negated	3	0	1
	Family (affirmed)	0	0	—
	Family (uncertain)	0	0	—
	Family (negated)	2	0	1
Overall				
	Affirmed	152	26	0.854
	Uncertain	34	7	0.829
	Negated	66	5	0.93
	All	252	38	0.87

6.4 Discussion

We have described the development and implemention of a patient identificiation and phenotyping pipeline, and applied it to the discovery and phenotyping of patients with HCM at UHB. The automated validation shows a high recall for HCM, AF, and HF against registry and ECG sources. The precision of HCM against the registry source is low, at 0.47, because the pipeline discovered more patients than exist in these databases. This is confirmed by the manual validation, which reveals high precision for the task of affirming the three conditions. With the assumption that this precision holds for the entire result set, the pipeline has effectively identified a large number of new patients currently

unknown to the rare disease registry or HES, and patients known to the registry but with important comorbidities unknown to it. These patients can be manually validated, and brought into care at the rare disease clinic at the hospital - most deaths caused by HCM are in the unmanaged population. In addition, prioritisation of patient treatment is performed in part on the basis of important complications such as AF and HF, as they vastly increase the chance of further complications and sudden death. Better knowledge of patients with these complications enables for prioritisation of high-risk patients who may have otherwise been missed.

Recall for the HES validation was lower than for the registry, and this is likely because only letters mentioning HCM were evaluated. It is, however, especially low for HF, and we wonder whether this is due to a coding error, or criteria for the use of the HF ICD code that differ from explicit diagnosis (there are differing definitions of what constitutes heart failure, for example, based on clinical presentation versus ejection fraction measurements). For future work, we would like to consider the entire patient record. The evaluation could have also been improved by matching document dates with the event dates provided in the HES database.

There are, however, limitations of the evaluation. During the training phase the irrelevancy vocabulary was expanded with phrases that the uncertainty and negation algorithms could not correctly classify. For example, the phrase 'ruled out' can't be captured, because it spans multiple words. By adding it to the irrelevancy vocabulary, sentences including such phrases were excluded from further classification. Because they are labelled, this is not a problem for the given recall or precision metrics (errors in this algorithm are counted in the overall metrics). However, because the pipeline is designed to use irrelevancy classification as a pre-processing step, 'irrelevant' sentences are deleted, and aren't considered in the manual validation. While this would not affect the precision of the given metrics, it means there is no evaluation of the precision of the irrelevancy classifier, affecting the unmeasured recall of the negation and uncertainty classifiers (patients who should be classified as uncertain or negated are considered irrelevant). Another problem

with the manual validation is that the correctness of classifications are binary, and no information is returned about what the correct classification is, or what the reason for the error is. For example, we cannot tell whether a false positive affirmation is because the sentences do not refer to the concept in question, or because they should be classified as uncertain or negated. This is also true for family classifications. As another example, we cannot tell whether the one false positive for affirmitive family history in HCM is because the sentences don't refer to a family member, don't refer to HCM, or because they say there is no family history. This kind of increased resolution in the evaluation would allow for better feedback for algorithm improvement during the training phase, and would also allow for the calculation of meaningful recall statistics for each classification. These problems were found during the evaluation stage of the project, and we plan to modify the pipeline and evaluation process to account for them.

Nevertheless, while the results of the manual validation cannot speak to the recall of the classifiers, they do show that sentence based disambiguative classifiers with a simple count-based decision procedure are highly precise for labelling the overall status of a patient's relationship with a condition. The manual precision for HCM is substantially lower than for AF and HF, which we expect is because unlike the other comorbidities, it is a complex disease with a complex presentation, and is therefore discussed in more depth and diagnosed less frequently. While all of the classifiers perform well, the uncertainty classifier has slightly lower precision. During the training phase, we noticed that many false positives were caused by the seed vocabulary (re-used from a standard vocabulary for topic modelling) being unsuited for medical language, because the clinical narrative explored used a lot of hedging. For example, the sentence "the test **suggests** the patient has atrial fibrillation," would be classified as uncertain, but in most cases the validating clinician would mark these patients as affirmed. While these cases were excluded in the training set, it's possible that this did not solve additional errors of the same kind in the test set. It's possible that creating a new uncertainty vocabulary from scratch may improve performance.

A clinical expert (WB) reviewed the false negative results of the pipeline with respect to the rare disease registry, and found that the majority of these patients were actually genotype positive, but negative for the phenotype of HCM. These patients are usually tested on account of a family history of having the disease, and then followed up every few years by the specialist clinic to monitor for development of any clinical presentation of the disease. There is some uncertainty as to whether these patients actually have HCM.

While we have identified greater recall for the two comorbidities amongst the rare disease registry for two comorbidities, the registry includes more than 170 variables describing social history, imaging, experimental results, comorbidities, and outcomes for patients. As future work, we would like to extend the pipeline to capture these additional phenotypes. While the system could easily be extended to capture additional qualitiative phenotypes, it would require additional work to capture quantitative ones. These include important measurements such as left atrial volume, weight, or heart reate. There are also practical problems with validating such a large number of phenotypes for a large number of patients.

The aim of this experiment was to identify the current, or most recent known status of a patient. While HCM is a life-long condition, AF, HF, and other conditions may come and go. The current pipeline could be used to determine the patient's status at a particular time by generating a corpus with a particular cut-off date. For future work, we would like to update the algorithm to determine a patient's status at different time points. This would offer additional information about a patient's history.

CHAPTER 7

COMPLICATION PREDICTION

7.1 Introduction

As discussed in the *Literature Review*, most patients with HCM do not develop complications, and their life expectancies are not seriously affected. The patients at risk of complications, therefore, must be prioritised by the clinics that manage their condition. This is currently performed by clinicians, and the ACCF/AHA guidelines recommend that patients under 60 should undergo "comprehensive clinical assessments on an annual basis for risk stratification and evolution of symptoms[53]."

Prevention of sudden cardiac death is the foremost consideration for HCM, and most work around outcome prediction has focused on it. In the introduction we discussed two such risk calculation models based on time-to-event modelling that are used in clinical practice. However, HCM may also cause a number of other complications. These complications affect quality of life and increase the likelihood of progression to end-stage symptoms, including death. Earlier we discussed that certain lifestyle and prophylactic interventions can prevent the development and progression of these complications. Therefore, prioritisation of HCM patients based not only on the likelihood of sudden death in the medium term, but on the likelihood of developing complications, could constitute an important development in care.

In this chapter, we consider the development of two prognostic models for the devel-

opment of complications over a three year period in HCM patients. To do this, we will use the text mining pipeline described in the previous chapter to overcome limitations in the data reported in the rare disease registry. By identifying additional complication outcomes and assigning them dates, we will enable the construction of multivariable Cox regression models to predict the likelihood for development of AF or HF in HCM patients. We further hypothesise that the integration of the rare disease registry with structured data from other modalities, will uncover informative relationships between complication development and additional variables not captured in the registry itself.

7.2 Materials & Methods

The experiment consists of combining clinical data from multiple sources, organised by patient visits, identifying outcomes through a combination of specialist registry information and NLP-mined concepts. Using HF and AF outcomes, and time until their occurrence as events, we then construct two different time-to-event models based on the clinical data recorded at the time of each patient's first visit, to identify predictors associated with later development of those complications. The overall process is summarised in Figure 7.1.

All pre-processing, data integration, and modelling was performed using R. In the context of the time-to-event analysis, 'survival probability' is used to refer to the probability of the patient not experiencing the event at the given time-point.

The data was formulated from a combination of multi-modal data acquired from the UHB secondary care trust. They are described in Table 7.1. Patients in these datasets were linked with a pseudo-anonymous identifier, and other identifying information such as names and addresses were not provided.

The rare disease registry contains two tables, one describing the patient (one row per patient), and the other describing visits (one row per visit). There were 3,301 visits for 1,043 patients.

A combination of variables from the registry and NLP complications were used to

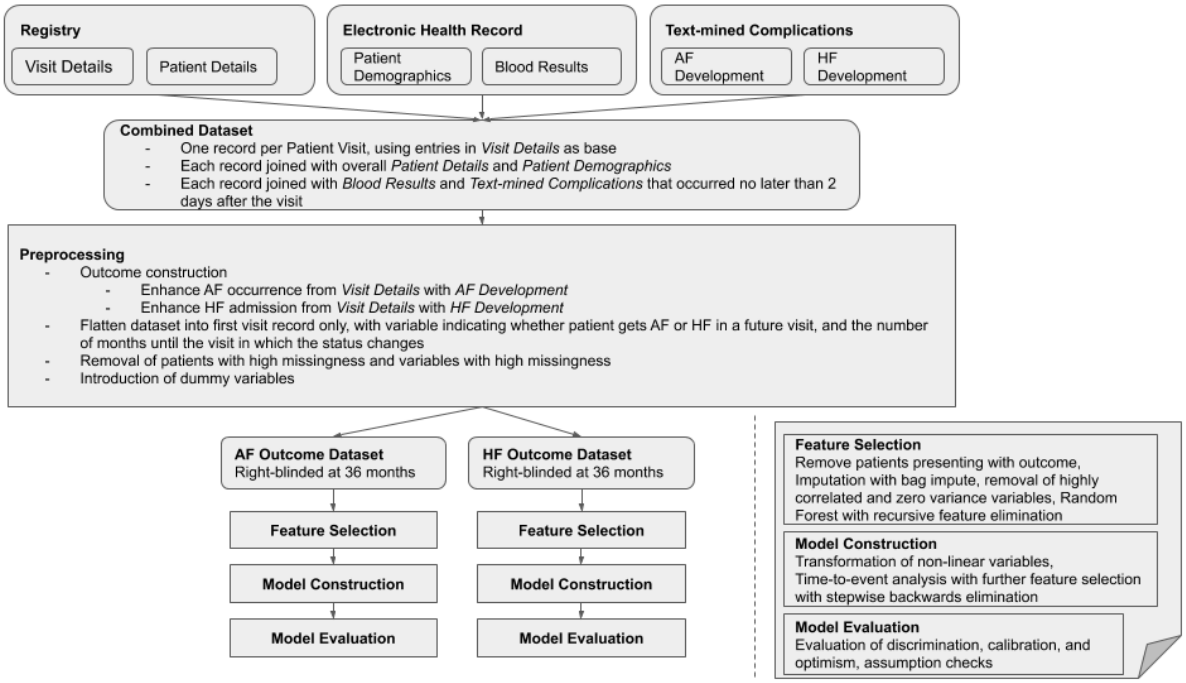

Figure 7.1: Summary of dataset construction and analysis pipeline. First, data from multiple sources are organised by patient visit. Outcome variables are then constructed using a combination of information recorded in the specialist registry and text-mined phenotypes. These values are then converted into time-to-event values, using 36 months as a cut-off, and only the first visit for each patient is then retained. After further pre-processing to remove variables and patients with high missingness, two separate datasets are created, one with AF as an outcome, and the other with a HF outcome; patients presenting with the relevant outcome are removed (as these patients cannot then undergo the event). These two datasets are then used to construct two separate time-to-event models, involving features selection, model construction, and model evaluation.

construct outcomes for analysis. The complications considered were atrial fibrillation and heart failure, because these were the comorbidities chosen for extraction in the previous chapter, on the basis of their importance in management of HCM. The registry records whether patients have AF, but it is missing cases (as was shown in the previous chapter), and is missing most information concerning when the patients developed it (making it unsuitable for time-to-event analysis). Its only information about HF is whether the patient has had a HF admission. While this is suitable to assume the patient has HF, it will not include all patients with HF, since many are managed through outpatient care only. Therefore, to construct our outcome variables, the AF and HF data in the visit table of the registry are extended with information from the text mining chapter

Table 7.1: Summary of data sources used for complication prediction modelling. All variables are considered as predictors, except for heart failure and atrial fibrillation, which is constructed through a combination of information from visit details and extracted complications (although heart failure is used as a predictor in the atrial fibrillation model, and atrial fibrillation is considered as a predictor in the heart failure model).

Name	Source	Description
Visit Details	Rare Disease Registry	Includes temporal comorbidity information, measurements and qualitative judgements on investigations performed on a particular visit to the clinic.
Patient Details	Rare Disease Registry	Patient background and comorbidity information particularly relevant to the condition.
Patient Demographics	EHR	Basic demographics and social history.
Blood Results	EHR	Continuous measurements from routine blood tests taken at the hospital.
Extracted Complications	NLP	Heart failure, atrial fibrillation, and when they were first diagnosed.

results. Diagnoses described by the NLP dataset are linked to the next visit recorded in the registry following it, updating that record to add complication cases where they are not present, and thereby providing a timeline for development of the complication.

These dates are then transformed into time-to-event values, measured in months, up to a maximum of three years (36 months), after which the outcome is right-blinded, and represents the number of months between the initial visit and the visit at which they developed the complication. This value was chosen on the basis of an evaluation of the distribution of months until complication development or final visit, which is shown in the results section. If they did not develop the complication, the time-to-event is the amount of months until their final visit. As described in the previous chapter, the NLP pipeline currently only uses the most recent HCM-related document mentioning the complication, and therefore time-to-event may be somewhat over-estimated in the model. While this is not optimal, we will find that most time-to-event values still fall within the 3 year period of interest, and this is because the rare disease registry has only been running for four years (although patients treated before it was established have been imported), and patients are usually followed up at intervals of at least a year. Therefore, we can treat the

NLP-derived outcomes in the same way as a blinded follow-up, which may often occur after the actual development of the clinical feature.

It is of note that many of the measurements could in some cases be considered time-dependent variables. For example, creatinine and other blood tests might normally be considered as time-dependent variables, especially in short term studies[161]. The visit details also encode information that changes over time. However, we expect that the proportional risk should stay constant with respect to the follow-up times considered in this study. We will examine the final model for adherence to the assumptions of the model. Creating a model with time-dependent variables could be considered as future work, in a more descriptive model that might be better suited for finding predictors that operate in a much shorter term or acute sense, or modelling responses to treatment.

The first visit was then selected for each patient, and then joined with the patient details table. These were then joined with the demographic information from the EHR, and also joined with the latest blood results that occurred at most two days after the date of the visit.

Patients who only had one visit were removed, as they could not be followed up. Variables with more than 50% missingness were removed, and patients with more missing variables than visits in the dataset were removed. Imputation was performed on the rest of the variables using bag impute, and highly correlated variables (above 0.75) were removed. Some approaches consider large amounts of missingness using multiple imputation, however removing predictors with a lot of missingness also helped for the purposes of reducing the dimensionality of the dataset. This led to 104 variables (including AF and HF), which are listed in Table 9.2 (in supplementary materials). When factors were flattened into dummy variables, this yielded 199 variables.

A dataset was then created for AF and HF separately. In the AF dataset, all patients who presented with AF were removed. For HF, all patients presenting with HF were removed. The relevant outcome was also removed as a predictor. This left 652 observations with 90 events for AF and 671 events with 35 events for HF. After this, variables

with no variance were removed, leading to 203 variables each. Missing variables were then imputed using bag impute. Variables with no variance were again removed, along with highly correlated variables above a cut-off of 0.75. This resulted in 187 variables considered for AF, and 124 variables considered for HF.

Feature selection was then performed using recursive feature elimination with a random forest model, after downsampling each dataset with respect to the outcome variable (events), and validating with 5-fold cross validation. We built a multivariable Cox regression model using the top 18 variables for AF, and the top 7 for HF, based on a relatively low events per variable setting of 5.

After the feature selection, we used general additive models to identify non-linear relationships. In the case of a non-linear relationship, we explored natural logarithms and fractional polynomial transformations for better representations of the variable. All models were fitted using the Cox proportional hazards model. This method was chosen due to its ability to create risk scoring models, and for its interpretability. It is also in common use for prognostic modelling in medicine, and therefore is relatively well understood within the clinical community.

For the multivariable model, backwards stepwise elimination was used to select the most informative variables for the final model. Backwards stepwise elimination starts with all variables selected in the model, iteratively removing those with the least information content, measuring whether or not its removal affects overall model performance beyond a certain threshold, adding it back if so. Schoenfeld residuals were used to check the proportional hazards assumption in the final model (these results are available in the supplementary material). Discrimination of the model is reported using concordance. Bootstrapping with n=1000 was used to estimate optimism, which we then use to provide an adjusted concordance. A calibration plot is provided, along with calibration slope statistic.

7.3 Results

7.3.1 Outcome Identification

Figure 7.2 shows the share of patients presenting with and developing complications before and after the dataset was extended with the NLP information. Only patients who did not present with the complication can be used, as patients who did present with the condition cannot then develop it. In the case of atrial fibrillation, 63 additional patients were found to present with the condition and 59 more developing the condition. In heart failure, 30 more were found presenting with the condition, and 36 more patients were found developing the condition. This shows that we have both reduced potential bias in the model by excluding patients who presented with the condition but were improperly reported in the registry, and found more developing events after the initial visit to be predicted. As we discovered in the previous chapter, the registry prevalence of AF is 21.65% (taken from undated entries associated with the patient rather than a visit, that were excluded in this investigation), and in the combined number of patients presenting with and developing AF, we have a prevalence of 20.23%, implying (with the high precision of the manual evaluation described by the previous chapter) that we have reasonably approximated the true situation.

After identifying the additional outcomes, we examined the distribution of time-to-event values (if not developing a complication, the time is the time until their final visit). These are shown in Figure 7.3. While the rare disease registry has only been running for four years, patients managed by the hospital previous to the registry have subsequently been imported. The figure also shows the time-to-event distribution after right-blinding to a period of 3 years (36 months).

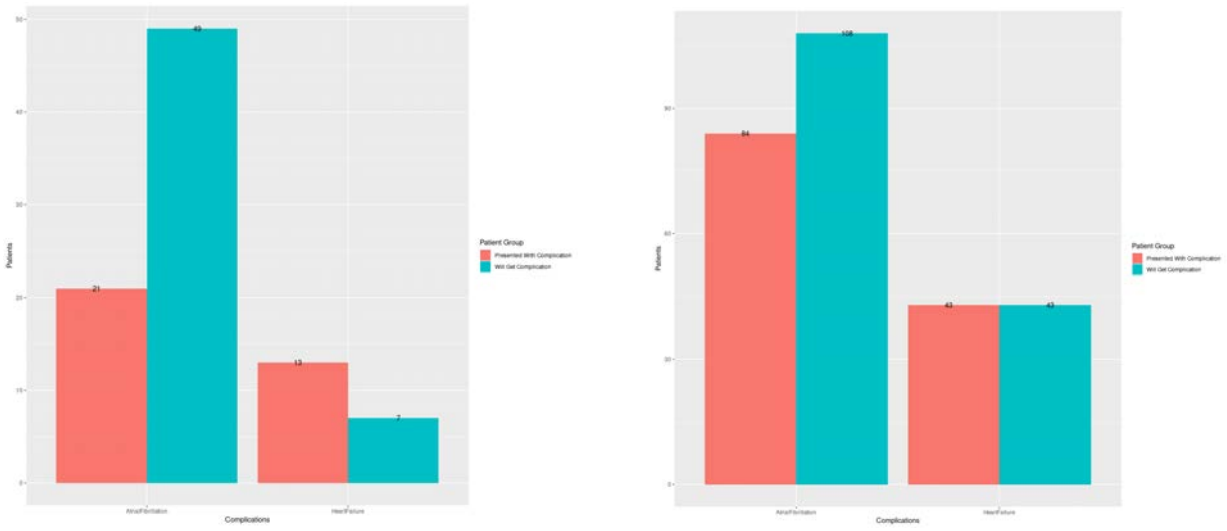

Figure 7.2: Comparison of number of patients presenting with and developing each complication both before extension with NLP information (left) and afterwards (right). The left graph shows the number of patients who developed AF or HF after their initial visit, according only to the visit details available in the structured data registry. On the right is the number of patients who developed HF and AF after their initial visit, when the registry data were extended with the AF and HF phenotype information mined from the clinical letters. The remainder of the experiments in this chapter use the constructed outcome variables represented by the right side of the figure.

7.3.2 Feature Selection

The results are summarised in Figure 7.4. Peak performance is found for atrial fibrillation at 100 variables, and heart failure at 50 variables. However, the number of events for atrial fibrillation and heart failure in our final dataset are 90 and 35, respectively. For the multivariable model, we aim to limit overfitting, and so limit the EPV to 5. Therefore, we will use the top 18 variables for the AF model, and 7 for the HF model. These variables are described in Table 9.3 (in supplementary materials).

7.3.3 Hazard Models

Atrial Fibrillation

The covariate values and p-values are summarised in Table 7.2. Backwards stepwise regression chose three variables for the model: age, LAVolume, and EGFR. We also

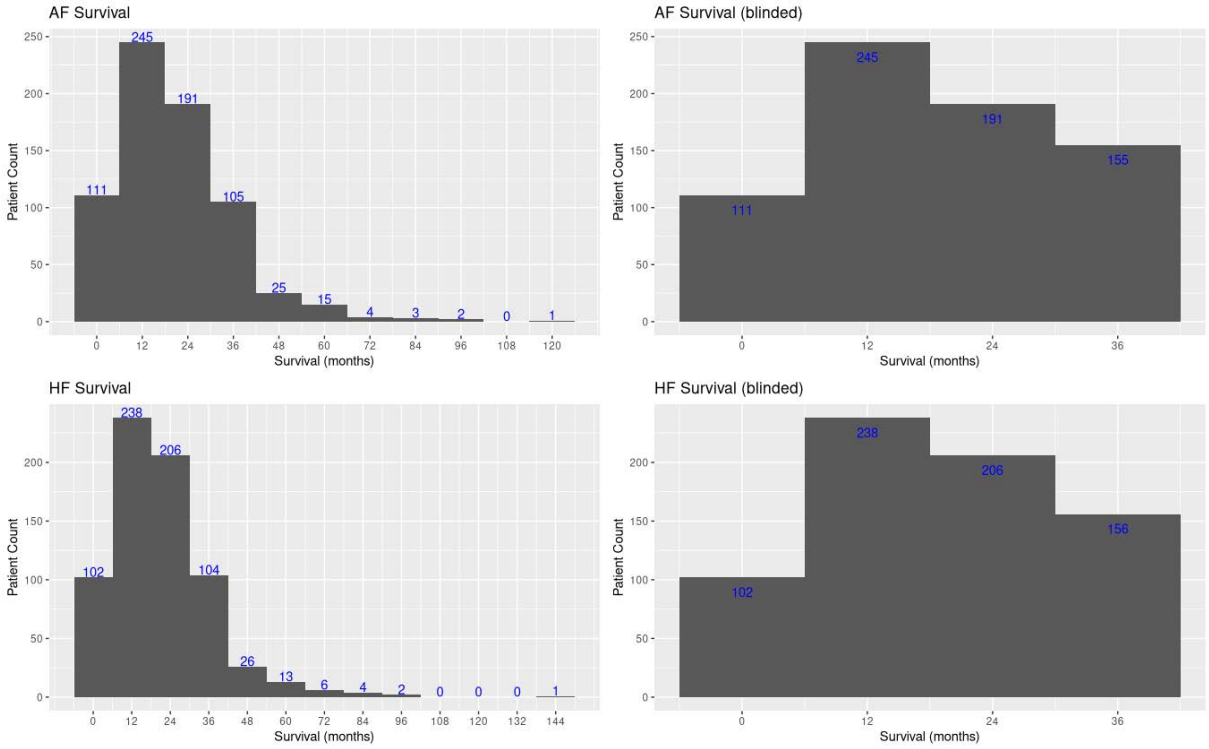

Figure 7.3: To identify a fixed time-to-event value to use for the study, we examined the distribution of months until either development of the outcome complication, or a patient's final visit to the specialist clinic, for both AF and HF: what can be considered the total duration of the study relevant to complication development for that patient. On the left, we see that the duration of study for the majority of patients falls within 36 months, or three years, for both outcomes. Therefore, we used 36 months as the maximal time point in the study, right-blinding patients beyond that point, meaning that patients who did not develop the relevant complication within 36 months, were recorded as not experiencing the event.

manually included BILI, because it was only removed by the stepwise regression on a borderline basis, and increased the discriminative power of the model. This indicates that the information criteria cut-off used as a parameter for the stepwise elimination was too coarse. EGFR is the most powerful independent predictor, but all predictors were highly significant independent predictors. Table 7.3 shows the evaluation metrics for the test, including discrimination and the likelihood ratio test, with applicable metrics adjusted for optimism through bootstrapping.

A calibration plot is shown in Figure 7.5. It shows some over-estimation of risk for high-risk patients, and some under-estimation of risk for medium-risk patients, but overall a good fit.

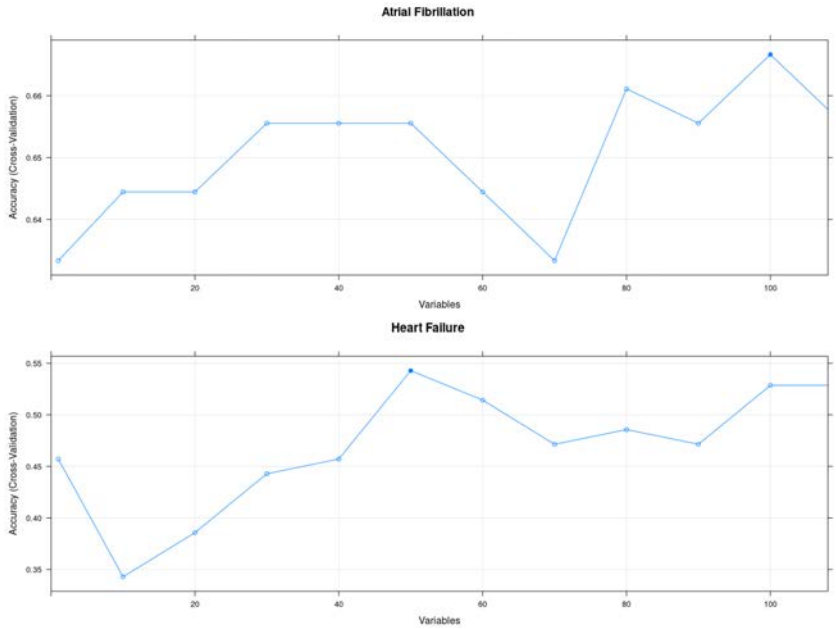

Figure 7.4: Plot of repeated feature elimination performance using Random Forest, for WillGetAtrialFibrillation and WillGetHeartFailure, measured by the accuracy statistic. Graph is cut off beyond 100 variables considered, since peak performance was found in the maximal case at 100.

Table 7.2: Multivariate coefficient values for atrial fibrillation model.

Name	Likelihood Ratio	.95 CI	p
Age	1.0402	1.0235-1.0571	1.65e-06
LAVolume	1.0152	1.0096-1.0208	1.08e-07
log(EGFR)	0.2945	0.1855-0.4676	2.19e-07
BILI	1.0491	1.0259-1.0728	2.67e-05

Table 7.3: Metrics for multivariable AF model

Name	Value	Optimism	Adjusted Value
Baseline Survival	0.8042181	—	—
c-index	0.799	0.00499	0.794
Calibration slope	1	0.1027	0.8973
Likelihood ratio	95.37 on 4 df, p=<2e-16	—	—

Heart Failure

The backwards stepwise regression for HF selected MCV and BNP. Due to the borderline exclusion of EOSINS and Age, we chose to manually include these variables in the prediction model, as manual evaluation indicated that overall model performance was negatively affected by the removal of these variables. This indicates that the information criteria

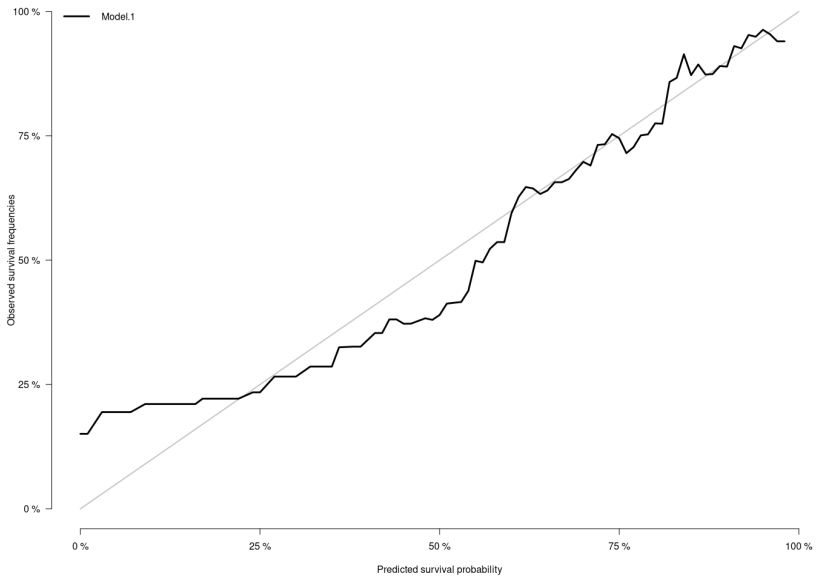

Figure 7.5: Calibration plot for the multivariable prediction model for development of AF.

cut-off used as a parameter for the stepwise elimination was too coarse. The model is summarised in Table 7.4, with metrics given in Table 7.5. There is some overfitting, as shown by the calibration slope shrinkage, but overall this did not make a substantial difference to the concordance of the model. The calibration plot for the HF model is shown in Figure 7.6. Since HF is a relatively rare complication, in comparison to AF, all survival probabilities are quite high. The fit overall is quite good, with some over-estimation of risk for patients with actual higher risk.

Table 7.4: Multivariate coefficient values for heart failure model.

Name	**Likelihood Ratio**	**.95 CI**	**p**
MCV	1.07305	1.004-1.1468	0.0375
EOSINS	0.01630	1.545e-05-17.1921	0.2464
log(Age)	2.21779	0.6524-7.5394	0.2020
I((BNP/1000)^1)	0.79380	0.6456-0.9761	0.0285

Table 7.5: Metrics for multivariable model for heart failure model

Name	Value	Optimism	Adjusted Value
Baseline Survival	0.9687578	—	—
c-index	0.723	0.022	0.701
Calibration slope	1	0.1689	0.8311
Likelihood ratio	27.14 on 4 df, p=1.862e-05	—	—

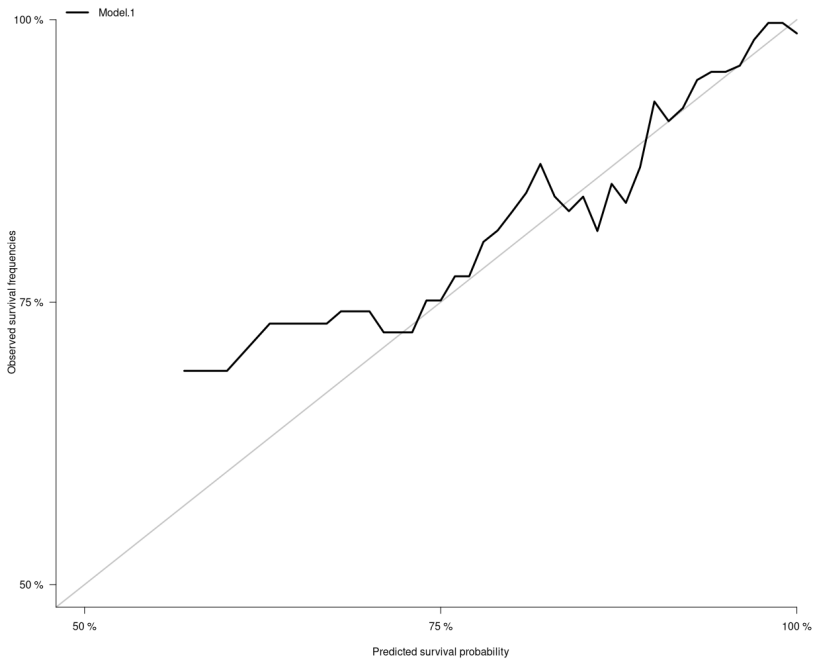

Figure 7.6: Calibration plot for the multivariable HF prediction model. Only values above 50% survival probability are shown, because all estimated and actual survival probabilities were above this.

7.4 Discussion

We have shown that using data extracted via our text mining pipeline, we can enable the development of predictive models that would not have been possible using the rare disease registry or structured EHR alone. Both models show a reasonable performance, with the discriminative power of the AF model being much higher. This is likely down to the fact that the incidence of HF was much lower, and therefore there was less data to train on. Both models were adjusted for optimism, and discrimination was not heavily affected by the shrinkage in either case. Investigation of the calibration plots showed that, in both cases, patients with a high actual risk had an over-estimated predicted risk. We believe

this to be an acceptable error in this context, because recall must be chosen over precision in detecting life-threatening complications.

The recursive feature elimination indicated that a higher performance may be achieved through the investigation of more variables. It's unknown whether this is due to overfitting of the feature selection, or whether there are other relationships. Without more observations, we cannot explore these additional variables without risking overfitting or multiple testing issues. However, investigation of more complicated variable interactions may have enabled the weaker predictors, especially those eliminated by the backward stepwise elimination to contribute more substantially to the predictions. Nevertheless, both models identified powerful independent predictors of complication development over a three year period.

Hijazi et al. reviewed known biomarkers in Atrial Fibrillation, listing low Glomerular Filtration Rate (eGFR/GFR) as being associated with prevalence of atrial fibrillation[70]. They also note a more general relationship between kidney function and AF, and with occurrence and outcomes of cardiovascular events. These facts may indicate that GFR may be also be an effective predictor in a generalised cardiovascular risk scoring system. Iguchi et al. explored the relationship between GFR and AF specifically, and found that prevalence of AF increases with a decreasing GFR[77]. Most importantly, the ESC guidlines for HCM note that renal function, particularly GFR and UR, "may be impaired in patients with severe left ventricular impairment." Their performance as predictors of AF complication in HCM has not previously been explored in the literature.

Olivotto et al. found that in HCM patients, AF was predicted by advancing age, and an increased LA volume at diagnosis[119]. While the patients in our study are usually not at the time point of their initial HCM diagnosis, (many patients will have previously been managed elsewhere), our work helps to confirm that these findings also hold for patients across the span of their disease. Tani et al. also found that left atrial volume was increased in HCM patients with paroxysmal hypertrophic cardiomyopathy[156]. Siontis Konstantinos C. et al. and Olivotto et al. note a significant relationship between age

and prevalence of AF in HCM patients[140]. Güngör et al. and Sarıkaya et al. found a relationship between red blood cell distribution width and AF. There is also some evidence of Bilirubin having some relationship with atrial fibrillation. Demir et al. found that bilirubin levels were significantly lower among patients with atrial fibrillation. The biological pathway for this is as yet unknown[47].

In terms of using these covariates as predictors, Benjamin found left ventricular hypertrophy (of which left atrial volume is an indicator) and age were independent risk factors for atrial fibrillation in a general population[26]. Losi et al. discuss predictors of AF development in HCM patients, finding that LA volume and age were co-dependent predictors of AF in HCM[97]. In summary, all of the variables in the model have known associations with AF in the general population, and most have also been explored in HCM populations. GFR and BILI however, have not been explored for prediction of future AF development in HCM patients or the general population, to our knowledge.

With respect to HF, BNP is considered to be an indicator, and is used by clinicians to manage the condition[39]. It has also been widely studied as a tool for patient stratification in HCM populations[39, 29]. Yang et al. discovered a relationship between RDW, a calculation derived from MCV, and hospital admissions for HF amongst HCM patients[174], and suggested that it might be a useful prognostic predictor. Several pieces of work have also identified relationships between RDW and HF diagnosis and prognosis in general populations[116, 167], as well as for general cardiac conditions[109]. However, as far as we know, it has not been studied as a predictor of the development of heart failure (although presumably borderline measurements are tracked). EOSINS has been used as a marker for particular kinds of heart failure[123], although other relationships seem unexplored. This could either be an artefact of overfitting or a novel relationship.

In both cases, our models determined significant discriminative and predictive relationships between predictors and development of the complication in a HCM population. In almost all cases, these associations have been discussed in the literature for HCM and general populations. However, several variables, most notably several of the routine

blood measurements, have not been previously considered in models predicting HF or AF in HCM populations. We think that this highlights a promising area for future exploration in the prediction of complications and patient prioritisation for HCM. Furthermore, since AF and HF both dramatically improve the risk of mortality in HCM, we think that these routine blood markers could also be considered for inclusion in models predicting sudden cardiac death.

As future work, we acknowledge the need to better characterise the cohort, and explore the potential relationships uncovered by this work. Alteration of the text mining pipeline to track the status of a condition in the text record over time would likely further improve upon the predictive power of models built using data from it, and would reduce potential bias from a lack of resolution on follow-up. From Figure 7.3, we also note that many follow ups occur within the first six months; it's possible that some patients with borderline or undiagnosed HF and AF (with according test results indicating so) are then diagnosed formally in the following months. This is not necessarily a problem, since we are predicting diagnosis of the complications, but could be a source of bias depending on the reporting criteria for the condition: perhaps the clinician suspects HF, but waits for the results of an MRI to make the formal diagnosis on the basis of an ejection fraction. This effect could potentially be confounded further by the NLP-derived data implying a later diagnosis date. It's also possible that some patients actually were diagnosed with AF or HF upon their initial visit, were misreported in the registry, and were then identified as developing it at a later date by the NLP algorithm. We do not believe that this can be too prevalent, however, as we would expect this to cause time-dependence and non-proportionality in the final model, as well as seriously affecting its discrimination.

Nevertheless, we believe that we have made an important initial step in characterising the dataset, and towards creating evidence-based tools that integrate data from several modalities to describe and potentially influence patient outcomes. Once the previously mentioned limitations are solved, we plan to construct a generalisable model to be considerd for external validation and potential use in clinical practice.

CHAPTER 8

CONCLUSIONS

In this thesis we have explored several fields in biomedical informatics, describing contributions to each. These fields and contributions are deeply inter-linked, and subsequent chapters build on the work described by previous chapters, either directly or indirectly. In the *unMIREOT* chapter, we surveyed the biomedical ontology ecosystem for interoperability issues, revealing extremely large clusters of hidden unsatisfiability. We provided a tool to diagnose and automatically repair these inconsistencies, applying it to the OBO Foundry to discover a small number of axioms accounting for all cases of inconsistency. The *Synonym Expansion* chapter presents a novel method of exploring ontology reuse and redundancy to obtain extended sets of synonyms for terms. The results of the synonym expansion method are affected by the quality of ontology interoperability, since hidden inconsistencies could potentially lead to imprecision at the equivalency step, returning synonyms from terms that are actually disjoint from the target. While the novel *Negation Detection* algorithm does not directly benefit from the ontology work, it does show superior performance over two clinical text corpora, and is used in combination with the *Synonym Expansion* work to form the basis of the text mining pipeline described by the *Patient Identification and Phenotype Extraction* chapter. The pipeline discovered diagnosed patients with HCM unknown to the rare disease registry, who will now be brought under specialist care. This, along with the better knowledge obtained of which patients suffer HCM complications, will directly influence patient care. In the following chap-

ter, the previous work culminates in these extracted phenotypes being used to enable an investigation of predictors of AF and HF in a HCM population. While not necessarily immediately generalisable, this work revealed promising relationships between several routine blood markers future complication development in a HCM cohort, which have not previously been explored in prognostic models.

Each of the earlier chapters also present individual contributions divorced from their context in the final two. The MIREOT investigation reveals problems and solutions that may affect any use of biomedical ontologies, and the two novel algorithms for text mining consist of components that can be used in any text mining pipeline or experiment. Meanwhile, the text mining pipeline itself can be applied to any condition or group of conditions, even outside of biomedicine.

Due to the scope of the work, and the limited time involved, we were unable to fully evaluate every piece of work with respect to how it relates to the other pieces of work. While work in individual chapters is evaluated, and the results are used in subsequent chapters, the contribution of the included work is not measured. For example, we have not proven that the synonym expansion module or negation detection module improve the results of the text mining pipeline. It is likely that since the performance is better than other algorithms in isolation, this translates to improved performance as part of an information extraction pipeline. However, this would ideally be proven by comparing the performance of an information extraction task using several negation modules. The same kind of investigation could also be used to test the efficacy of the synonym expansion algorithm. While we expect that the ontology interoperability issues revealed in the *unMIREOT* chapter has implications for the synonym expansion algorithm, we did not test this, as it would ultimately require the ontology developers to repair the root causes of the inconsistencies. It would be possible to test whether ontologies repaired by the automatic repair algorithm produce better results for vocabulary expansion by creating a local staging version of AberOWL. We would also like to explore whether such changes improve the performance of semantic similarity experiments, for example in predicting

patient-patient similarity.

We also identified several limitations of each individual chapter, and areas we would like to explore in future work. The most important overall limitation is that the phenotyping pipeline only classifies the most recent document discussing the condition, as this limits the resolution and potentially the correctness of the prognostic models at intermediate time points. Upon extending the pipeline with this functionality, we would like to build a new model that can be published and externally validated.

We would also like to explore other uses of the text extraction pipeline, especially as it applies to analysis. Text mining of literature reviews could be used to perform pre-feature selection for prognostic models. Ontology-based integration could also enable the use of public data or different sources of structured data to be used in modelling. We would also like to explore the use of ontology-based analysis, such as semantic similarity, for prediction. Furthermore, this thesis has focused on moving from general background knowledge expressed by ontologies, to extracting more specific knowledge from text, to then exploring an even more specific development of complications. However, we also expect that text mining can be used to move in the other direction, and we would like to investigate using some of the technologies developed in this thesis to move from descriptive text, such as clinical letters, to extending existing background knowledge and perhaps even creating new ontologies from the knowledge expressed in text.

CHAPTER 9

SUPPLEMENTARY MATERIALS

In the *Term Expansion*, *Negation Detection*, and *Patient Identification and Extraction* chapters, we refer to a clinical expert WB. This is Dr William Bradlow, a consultant cardiologist at UHB.

9.1 Negation Detection

Table 9.1: Summary the negation algorithms compared with in the negation detection algorithm, including versions and source for download.

Name	Version	Source
NegEx	Commit 21b013c	`https://github.com/chapmanbe/negex/`
pyConTextNLP	0.7.0.1	`https://pypi.org/project/pyConTextNLP/`
negation-detection	Commit 6d9d88e	`https://github.com/gkotsis/negation-detection`
NegBio	Commit d025875	`https://github.com/ncbi-nlp/NegBio/`

9.2 Risk Prediction

Table 9.2: Summary of variables considered after initial pre-processing. Note that atrial fibrillation was not considered for experiments with atrial fibrillation as an outcome, and heart failure was not considered for experiments with heart failure as an outcome.

Category	**Variables**
Blood Results	ALB, ALP1, ALT, BASOS, BILI, CK, CREAT, EGFR, EOSINS, FT4, GFR, HCT, HGB, K, LYMPHS, MCH, MCHC1, MCV, MONOS, NA, NEUTS, PLATS, RBC, RDW, TSH, UR, WBC
Family, Social History, and Demographic	Age, BMI, Female, AlcoholIntake, ExerciseIntensity, SmokingHistory, SmokingPackYears, MonthsSinceDiagnosis, FamilyAtrialFibrillation, FamilyHCM, FamilyHeartFailure, AgeAtDiagnosis
Routinely collected healthcare data	BloodPressureTotal, BloodPressureAvgBPDia, BloodPressureAvgBPSys, BNPngL, BNPngLAvgVal
Comorbidities	AbortedSuddenCardiacDeath, AcuteKidneyInjury, AlcoholSeptalAblation, AnginaPectoris, ApicalVariantphenotypepresence, AtrialFibrillation, HeartFailure, BasalPhenotype, CABG, ChronicKidneyDiseases, COPD, CoronaryArteryDisease, CRT, Depression, DiabetesMelitusTypeII, ObstructiveSleepApnoea, Oedema, Orthopnea, Defibrillator, Palpitations, ParoxysmalDyspnea, PercutaneousCoronaryIntervention , NonAnginalAtypicalChestPain, PPM, Stroke, TIA, UnexplainedSyncope, HistoryOfArrythmia, Hypercholesterolemia, Hypertension, ICD, Lethargy, MitralValveSurgery, Myectomy, MyocardialInfarction
HCM-related Phenotypes	NYHA, SCD, NonObstructionPhenotype, SigmoidSeptalPhenotype, LabilePhenotype, NeutralSeptalPhenotype
Genetic	FabryEnzymeLevel, MYBPC3Other, MYBPC3PM, MYBPC3VUS, MYH7Other, MYH7PM, MYH7VUS, TNNI3Other, TNNI3PM, TNNT2Other, TNNT2PM, TNNT2VUS
ECG	ECGLBBB, ECGPacedRhythm, ECGPreexcitation, ECGRBBB, ECGRhythm

Table 9.3: Summary of variables selected by the feature selection algorithm for consideration in the AF and HF complication prediction models.

Name	Model	Description	Unit
LA volume	AF	Volume of the left atrium. Measured either by MRI or Echocardiogram, with MRI preferred (as it is more accurate).	mL
TNNI3Other	AF	An ambiguous result for the TNNI3 pathogenic mutation.	Boolean
EGFR	AF	Estimated Glomerular Filtration Rate.	mL/min.
GFR	AF	Glomerular Filtration Rate.	mL/min
Age at diagnosis	AF	Age of the patient at the time they were diagnosed.	Integer years
MYH7PM	AF	Whether the patient has the MYH7 pathogenic mutation.	Boolean
Months since diagnosis	AF	Number of months since the patient was first diagnosed with HCM.	Integer months
MWTSeptum	AF	The thickest wall measurement of the left ventricular septum.	mm
BASOS	AF	Absolute basophil count in blood.	$10^3/\mu$L
LYMPHS	AF	Absolute lymphocyte count in blood.	10^9/L
BILI	AF	Bilirubin level in the blood.	mmol/L
Coronary artery disease	AF	Whether the patient is diagnosed with coronary artery disease.	Factor: Never, Present, Previous
RDW	AF	Red Cell Distribution Width in blood.	Percentage
MCV	AF, HF	Mean Corpuscular Volume in blood.	fL
EOSINS	AF, HF	Eosinophil count in blood.	10^9/L
Age	AF, HF	Patient's age at the time of the visit.	Integer years
BNP	AF, HF	Level of Brain Natriuretic Peptide (BNP) hormone in the blood.	pmol/L.
UR	AF, HF	Hematuria (blood in urine) test.	red blood cells/high-power field
ALP1	HF	Alkaline Phosphatas Level blood test.	U/L
NA	HF	Blood sodium measurement.	mmol/L

Table 9.4: Test of proportional hazards assumption for final AF complication prediction model via Schoenfield residuals.

Variable	Chi Sq	df	p
Age	1.1080	1	0.29
LAVolume	0.1140	1	0.74
log(EGFR)	0.0954	1	0.76
BILI	1.5347	1	0.22
Overall	3.3426	4	0.50

Table 9.5: Test of proportional hazards assumption for final HF complication prediction model via Schoenfield residuals.

Variable	**Chi Sq**	**df**	**p**
MCV	0.193	1	0.66
EOSINS	0.684	1	0.41
log(Age)	0.208	1	0.65
I((BNP/1000)$^{-}$1)	0.699	1	0.40
GLOBAL	1.592	4	0.81

LIST OF REFERENCES

[1] DCMI: DCMI Metadata Terms. https://www.dublincore.org/specifications/dublin-core/dcmi-terms/.

[2] DISEASE — meaning in the Cambridge English Dictionary. https://dictionary.cambridge.org/dictionary/english/disease.

[3] Disease. https://www.wikidata.org/wiki/Q12136.

[4] About: Heart failure. http://dbpedia.org/page/Heart_failure.

[5] Hypertrophic cardiomyopathy (HCM). https://www.cardiomyopathy.org/hypertrophic-cardiomyopathy/intro.

[6] ICD-9-CM Diagnostic Codes to SNOMED CT Map. https://www.nlm.nih.gov/research/umls/mapping_projects/icd9cm_to_snomedct.html.

[7] MEDLINE®: Description of the Database. https://www.nlm.nih.gov/bsd/medline.html.

[8] Ontology Alignment Evaluation Initiative::Home. http://oaei.ontologymatching.org/.

[9] RDF Schema 1.1. https://www.w3.org/TR/rdf-schema/, .

[10] RDF Schema 1.1. https://www.w3.org/TR/rdf-schema/#ch_label, .

[11] SKOS Core Vocabulary Specification. https://www.w3.org/TR/swbp-skos-core-spec/#altLabel.

[12] STATO: An Ontology of Statistical Methods. http://stato-ontology.org/.

[13] 2014 ESC Guidelines on diagnosis and management of hypertrophic cardiomyopathy: The Task Force for the Diagnosis and Management of Hypertrophic Cardiomyopathy of the European Society of Cardiology (ESC). *European Heart Journal*, 35(39):2733–2779, October 2014. ISSN 0195-668X, 1522-9645. doi: 10.1093/eurheartj/ehu284.

[14] Information-artifact-ontology/IAO. IAO, October 2019.

[15] Heart failure. *Wikipedia*, January 2020.

[16] Mehdi Allahyari, Seyedamin Pouriyeh, Mehdi Assefi, Saied Safaei, Elizabeth D. Trippe, Juan B. Gutierrez, and Krys Kochut. A Brief Survey of Text Mining: Classification, Clustering and Extraction Techniques. July 2017.

[17] Douglas G. Altman and Patrick Royston. The cost of dichotomising continuous variables. *BMJ*, 332(7549):1080, May 2006. ISSN 0959-8138, 1468-5833. doi: 10.1136/bmj.332.7549.1080.

[18] Douglas E. Appelt. Introduction to information extraction. *AI Communications*, 12(3):161–172, August 1999. ISSN 0921-7126.

[19] Aristotle and Porphyry. *The Organon, Or Logical Treatises, of Aristotle: With the Introduction of Porphyry. Literally Translated, with Notes, Syllogistic Examples, Analysis, and Introduction.* H.G. Bohn, 1853.

[20] Michael Ashburner, Catherine A. Ball, Judith A. Blake, David Botstein, Heather Butler, J. Michael Cherry, Allan P. Davis, Kara Dolinski, Selina S. Dwight, Janan T. Eppig, Midori A. Harris, David P. Hill, Laurie Issel-Tarver, Andrew Kasarskis, Suzanna Lewis, John C. Matese, Joel E. Richardson, Martin Ringwald, Gerald M. Rubin, and Gavin Sherlock. Gene Ontology: Tool for the unification of biology.

Nature Genetics, 25(1):25–29, May 2000. ISSN 1061-4036, 1546-1718. doi: 10.1038/75556.

[21] Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary Ives. DBpedia: A Nucleus for a Web of Open Data. In Karl Aberer, Key-Sun Choi, Natasha Noy, Dean Allemang, Kyung-Il Lee, Lyndon Nixon, Jennifer Golbeck, Peter Mika, Diana Maynard, Riichiro Mizoguchi, Guus Schreiber, and Philippe Cudré-Mauroux, editors, *The Semantic Web*, Lecture Notes in Computer Science, pages 722–735, Berlin, Heidelberg, 2007. Springer. ISBN 978-3-540-76298-0. doi: 10.1007/978-3-540-76298-0_52.

[22] Anita Bandrowski, Ryan Brinkman, Mathias Brochhausen, Matthew H. Brush, Bill Bug, Marcus C. Chibucos, Kevin Clancy, Mélanie Courtot, Dirk Derom, Michel Dumontier, Liju Fan, Jennifer Fostel, Gilberto Fragoso, Frank Gibson, Alejandra Gonzalez-Beltran, Melissa A. Haendel, Yongqun He, Mervi Heiskanen, Tina Hernandez-Boussard, Mark Jensen, Yu Lin, Allyson L. Lister, Phillip Lord, James Malone, Elisabetta Manduchi, Monnie McGee, Norman Morrison, James A. Overton, Helen Parkinson, Bjoern Peters, Philippe Rocca-Serra, Alan Ruttenberg, Susanna-Assunta Sansone, Richard H. Scheuermann, Daniel Schober, Barry Smith, Larisa N. Soldatova, Christian J. Stoeckert, Chris F. Taylor, Carlo Torniai, Jessica A. Turner, Randi Vita, Patricia L. Whetzel, and Jie Zheng. The Ontology for Biomedical Investigations. *PLoS ONE*, 11(4), April 2016. ISSN 1932-6203. doi: 10.1371/journal.pone.0154556.

[23] Riccardo Bellazzi and Blaz Zupan. Predictive data mining in clinical medicine: Current issues and guidelines. *International journal of medical informatics*, 77: 81–97, March 2008. doi: 10.1016/j.ijmedinf.2006.11.006.

[24] Ofir Ben-Assuli. Electronic health records, adoption, quality of care, legal and privacy issues and their implementation in emergency departments. *Health Policy*, 119(3):287–297, March 2015. ISSN 0168-8510. doi: 10.1016/j.healthpol.2014.11.014.

[25] Ralf Bender and Ulrich Grouven. Logistic regression models used in medical research are poorly presented. *BMJ*, 313(7057):628, September 1996. ISSN 0959-8138, 1468-5833. doi: 10.1136/bmj.313.7057.628.

[26] Emelia J. Benjamin. Independent Risk Factors for Atrial Fibrillation in a Population-Based Cohort: The Framingham Heart Study. *JAMA*, 271(11):840, March 1994. ISSN 0098-7484. doi: 10.1001/jama.1994.03510350050036.

[27] Tim Benson. The history of the Read Codes: The inaugural James Read Memorial Lecture 2011. *Informatics in Primary Care*, 19(3):173–182, 2011. ISSN 1476-0320. doi: 10.14236/jhi.v19i3.811.

[28] TIM BERNERS-LEE, JAMES HENDLER, and ORA LASSILA. THE SEMANTIC WEB. *Scientific American*, 284(5):34–43, 2001. ISSN 0036-8733.

[29] Josepha Binder, Steve R. Ommen, Horng H. Chen, Michael J. Ackerman, A. Jamil Tajik, and Allan S. Jaffe. Usefulness of Brain Natriuretic Peptide Levels in the Clinical Evaluation of Patients With Hypertrophic Cardiomyopathy. *The American Journal of Cardiology*, 100(4):712–714, August 2007. ISSN 0002-9149. doi: 10.1016/j.amjcard.2007.03.089.

[30] Christopher Brewster, Harith Alani, Srinandan Dasmahapatra, and Yorick Wilks. Data Driven Ontology Evaluation. In *International Conference on Language Resources and Evaluation (30/05/04)*, 2004.

[31] Gavin Carothers and Eric Prud'hommeaux. RDF 1.1 turtle. W3C recommendation, W3C, February 2014.

[32] W. W. Chapman, W. Bridewell, P. Hanbury, G. F. Cooper, and B. G. Buchanan. A simple algorithm for identifying negated findings and diseases in discharge summaries. *Journal of Biomedical Informatics*, 34(5):301–310, October 2001. ISSN 1532-0464. doi: 10.1006/jbin.2001.1029.

[33] Yoo Jin Cho, James F. Thrasher, Kamala Swayampakala, Isaac Lipkus, David Hammond, Kenneth Michael Cummings, Ron Borland, Hua-Hie Yong, and James W. Hardin. Does Adding Information on Toxic Constituents to Cigarette Pack Warnings Increase Smokers' Perceptions About the Health Risks of Smoking? A Longitudinal Study in Australia, Canada, Mexico, and the United States. *Health Education & Behavior: The Official Publication of the Society for Public Health Education*, 45 (1):32–42, February 2018. ISSN 1552-6127. doi: 10.1177/1090198117709884.

[34] Evangelia Christodoulou, Jie Ma, Gary S. Collins, Ewout W. Steyerberg, Jan Y. Verbakel, and Ben Van Calster. A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. *Journal of Clinical Epidemiology*, 110:12–22, June 2019. ISSN 0895-4356. doi: 10.1016/j.jclinepi.2019.02.004.

[35] K. Bretonnel Cohen, Martha Palmer, and Lawrence Hunter. Nominalization and Alternations in Biomedical Language. *PLOS ONE*, 3(9):e3158, September 2008. ISSN 1932-6203. doi: 10.1371/journal.pone.0003158.

[36] Gary S. Collins, Johannes B. Reitsma, Douglas G. Altman, and Karel G. M. Moons. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): The TRIPOD statement. *BMJ (Clinical research ed.)*, 350: g7594, January 2015. ISSN 1756-1833. doi: 10.1136/bmj.g7594.

[37] Ronald Cornet and Nicolette de Keizer. Forty years of SNOMED: A literature review. *BMC Medical Informatics and Decision Making*, 8(1):S2, October 2008. ISSN 1472-6947. doi: 10.1186/1472-6947-8-S1-S2.

[38] Mélanie Courtot, Frank Gibson, Allyson L. Lister, James Malone, Daniel Schober, Ryan R. Brinkman, and Alan Ruttenberg. MIREOT: The minimum information to reference an external ontology term. *Applied Ontology*, 6(1):23–33, January 2011. ISSN 1570-5838. doi: 10.3233/AO-2011-0087.

[39] Martin R Cowie and Gustavo F Mendez. BNP and congestive heart failure. *Progress in Cardiovascular Diseases*, 44(4):293–321, January 2002. ISSN 0033-0620. doi: 10.1053/pcad.2002.24599.

[40] D. R. Cox. Regression Models and Life-Tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 34(2):187–202, 1972. ISSN 2517-6161. doi: 10.1111/j.2517-6161.1972.tb00899.x.

[41] Isabel F. Cruz, Flavio Palandri Antonelli, and Cosmin Stroe. AgreementMaker: Efficient matching for large real-world schemas and ontologies. *Proceedings of the VLDB Endowment*, 2(2):1586–1589, August 2009. ISSN 2150-8097. doi: 10.14778/1687553.1687598.

[42] Hamish Cunningham. GATE, a General Architecture for Text Engineering. *Computers and the Humanities*, 36(2):223–254, May 2002. ISSN 1572-8412. doi: 10.1023/A:1014348124664.

[43] 'Rafal Dabrowski'. Use of antiplatelet and anticoagulant drugs in hypertension. https://www.escardio.org/Journals/E-Journal-of-Cardiology-Practice/Volume-13/use-of-antiplatelet-and-anticoagulant-drugs-in-hypertension, https://www.escardio.org/Journals/E-Journal-of-Cardiology-Practice/Volume-13/use-of-antiplatelet-and-anticoagulant-drugs-in-hypertension.

[44] John Day-Richter, Midori A. Harris, Melissa Haendel, and Suzanna Lewis. OBO-Edit—an ontology editor for biologists. *Bioinformatics*, 23(16):2198–2200, August 2007. ISSN 1367-4803. doi: 10.1093/bioinformatics/btm112.

[45] Paula de Matos, Adriano Dekker, Marcus Ennis, Janna Hastings, Kenneth Haug, Steve Turner, and Christoph Steinbeck. ChEBI: A chemistry ontology and database. *Journal of cheminformatics*, 2(S1):P6, 2010.

[46] Chiara Del Vescovo, Damian D. G. Gessler, Pavel Klinov, Bijan Parsia, Ulrike Sattler, Thomas Schneider, and Andrew Winget. Decomposition and Modular

Structure of BioPortal Ontologies. In Lora Aroyo, Chris Welty, Harith Alani, Jamie Taylor, Abraham Bernstein, Lalana Kagal, Natasha Noy, and Eva Blomqvist, editors, *The Semantic Web – ISWC 2011*, Lecture Notes in Computer Science, pages 130–145, Berlin, Heidelberg, 2011. Springer. ISBN 978-3-642-25073-6. doi: 10.1007/978-3-642-25073-6_9.

[47] Mehmet Demir, Canan Demir, Umut Uyan, and Mehmet Melek. The Relationship Between Serum Bilirubin Concentration and Atrial Fibrillation. *Cardiology Research*, 4(6):186–191, December 2013. ISSN 1923-2829. doi: 10.4021/cr299w.

[48] Jürgen Dönitz and Edgar Wingender. The ontology-based answers (OBA) service: A connector for embedded usage of ontologies in applications. *Frontiers in genetics*, 3:197, 2012.

[49] Christopher Funk, William Baumgartner, Benjamin Garcia, Christophe Roeder, Michael Bada, K Bretonnel Cohen, Lawrence E Hunter, and Karin Verspoor. Large-scale biomedical concept recognition: An evaluation of current automatic annotators and their parameters. *BMC Bioinformatics*, 15:59, February 2014. ISSN 1471-2105. doi: 10.1186/1471-2105-15-59.

[50] Christopher S. Funk, K. Bretonnel Cohen, Lawrence E. Hunter, and Karin M. Verspoor. Gene Ontology synonym generation rules lead to increased performance in biomedical concept recognition. *Journal of Biomedical Semantics*, 7(1):52, September 2016. ISSN 2041-1480. doi: 10.1186/s13326-016-0096-7.

[51] Dov M. Gabbay and Philippe Smets, editors. *Handbook of Defeasible Reasoning and Uncertainty Management Systems: Algorithms for Uncertainty and Defeasible Reasoning*. Handbook of Defeasible Reasoning and Uncertainty Management Systems. Springer Netherlands, 2000. ISBN 978-0-7923-6672-0. doi: 10.1007/978-94-017-1737-3.

[52] Matthias Ganzinger, Shan He, Kai Breuhahn, and Petra Knaup. On the ontology

based representation of cell lines. *PloS One*, 7(11):e48584, 2012. ISSN 1932-6203. doi: 10.1371/journal.pone.0048584.

[53] Gersh Bernard J., Maron Barry J., Bonow Robert O., Dearani Joseph A., Fifer Michael A., Link Mark S., Naidu Srihari S., Nishimura Rick A., Ommen Steve R., Rakowski Harry, Seidman Christine E., Towbin Jeffrey A., Udelson James E., and Yancy Clyde W. 2011 ACCF/AHA Guideline for the Diagnosis and Treatment of Hypertrophic Cardiomyopathy. *Circulation*, 124(24):e783–e831, December 2011. doi: 10.1161/CIR.0b013e318223e2bd.

[54] Jeff Gilchrist, Monique Frize, Erika Bariciak, and Daphne Townsend. Integration of new technology in a legacy system for collecting medical data - challenges and lessons learned. *Conference proceedings: ... Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Annual Conference*, 2008:4326–4329, 2008. ISSN 1557-170X. doi: 10.1109/IEMBS.2008.4650167.

[55] George Gkotsis, Sumithra Velupillai, Anika Oellrich, Harry Dean, Maria Liakata, and Rina Dutta. Don't Let Notes Be Misunderstood: A Negation Detection Method for Assessing Risk of Suicide in Mental Health Records. In *Proceedings of the Third Workshop on Computational Linguistics and Clinical Psychology*, pages 95–105, San Diego, CA, USA, June 2016. Association for Computational Linguistics. doi: 10.18653/v1/W16-0310.

[56] Georgios V. Gkoutos, Chris Mungall, Sandra Dolken, Michael Ashburner, Suzanna Lewis, John Hancock, Paul Schofield, Sebastian Kohler, and Peter N. Robinson. Entity/quality-based logical definitions for the human skeletal phenome using PATO. In *2009 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 7069–7072, September 2009. doi: 10.1109/IEMBS.2009.5333362.

[57] Georgios V. Gkoutos, Paul N. Schofield, and Robert Hoehndorf. The neurobehavior ontology: An ontology for annotation and integration of behavior and behavioral phenotypes. In *International Review of Neurobiology*, volume 103, pages 69–87. Elsevier, 2012.

[58] Georgios V. Gkoutos, Paul N. Schofield, and Robert Hoehndorf. The Units Ontology: A tool for integrating units of measurement in science. *Database*, 2012, January 2012. doi: 10.1093/database/bas033.

[59] Sergey Goryachev. Implementation and Evaluation of Four Different Methods of Negation Detection. page 7.

[60] Tudor Groza, Sebastian Köhler, Sandra Doelken, Nigel Collier, Anika Oellrich, Damian Smedley, Francisco M Couto, Gareth Baynam, Andreas Zankl, and Peter N. Robinson. Automatic concept recognition using the Human Phenotype Ontology reference and test suite corpora. *Database: The Journal of Biological Databases and Curation*, 2015, February 2015. ISSN 1758-0463. doi: 10.1093/database/bav005.

[61] Barış Güngör, Kazım Serhan Özcan, İzzet Erdinler, Ahmet Ekmekçi, Ahmet Taha Alper, Damirbek Osmonov, Nazmi Çalık, Sukru Akyuz, Ercan Toprak, Hale Yılmaz, Aydın Yıldırım, and Osman Bolca. Elevated levels of RDW is associated with non-valvular atrial fibrillation. *Journal of Thrombosis and Thrombolysis*, 37(4):404–410, May 2014. ISSN 0929-5305, 1573-742X. doi: 10.1007/s11239-013-0957-1.

[62] Carol M. Hamilton, Lisa C. Strader, Joseph G. Pratt, Deborah Maiese, Tabitha Hendershot, Richard K. Kwok, Jane A. Hammond, Wayne Huggins, Dean Jackman, Huaqin Pan, Destiney S. Nettles, Terri H. Beaty, Lindsay A. Farrer, Peter Kraft, Mary L. Marazita, Jose M. Ordovas, Carlos N. Pato, Margaret R. Spitz, Diane Wagener, Michelle Williams, Heather A. Junkins, William R. Harlan, Erin M. Ramos, and Jonathan Haines. The PhenX Toolkit: Get the Most From Your Measures.

American Journal of Epidemiology, 174(3):253–260, August 2011. ISSN 0002-9262. doi: 10.1093/aje/kwr193.

[63] Stephen Hansen, Michael McMahon, and Andrea Prat. Transparency and Deliberation Within the FOMC: A Computational Linguistics Approach. *The Quarterly Journal of Economics*, 133(2):801–870, May 2018. ISSN 0033-5533. doi: 10.1093/qje/qjx045.

[64] Henk Harkema, John N. Dowling, Tyler Thornblade, and Wendy W. Chapman. Context: An Algorithm for Determining Negation, Experiencer, and Temporal Status from Clinical Reports. *Journal of biomedical informatics*, 42(5):839–851, October 2009. ISSN 1532-0464. doi: 10.1016/j.jbi.2009.05.002.

[65] Janna Hastings, Nina Jeliazkova, Gareth Owen, Georgia Tsiliki, Cristian R. Munteanu, Christoph Steinbeck, and Egon Willighagen. eNanoMapper: Harnessing ontologies to enable data integration for nanomaterial risk assessment. *Journal of Biomedical Semantics*, 6(1):10, March 2015. ISSN 2041-1480. doi: 10.1186/s13326-015-0005-5.

[66] Jill A. Hayden, Danielle A. van der Windt, Jennifer L. Cartwright, Pierre Côté, and Claire Bombardier. Assessing Bias in Studies of Prognostic Factors. *Annals of Internal Medicine*, 158(4):280, February 2013. ISSN 0003-4819. doi: 10.7326/0003-4819-158-4-201302190-00009.

[67] Yongqun He, Sirarat Sarntivijai, Yu Lin, Zuoshuang Xiang, Abra Guo, Shelley Zhang, Desikan Jagannathan, Luca Toldo, Cui Tao, and Barry Smith. OAE: The ontology of adverse events. *Journal of biomedical semantics*, 5(1):29, 2014.

[68] Pauline Heus, Johanna A. A. G. Damen, Romin Pajouheshnia, Rob J. P. M. Scholten, Johannes B. Reitsma, Gary S. Collins, Douglas G. Altman, Karel G. M. Moons, and Lotty Hooft. Poor reporting of multivariable prediction model stud-

ies: Towards a targeted implementation strategy of the TRIPOD statement. *BMC Medicine*, 16(1):120, July 2018. ISSN 1741-7015. doi: 10.1186/s12916-018-1099-2.

[69] Amanda Hicks, Mark A. Miller, Christian Stoeckert, and Danielle Mowery. The Hypertension Ontology. March 2019. doi: 10.5281/zenodo.2605329.

[70] Z. Hijazi, J. Oldgren, A. Siegbahn, C. B. Granger, and L. Wallentin. Biomarkers in atrial fibrillation: A clinical review. *European Heart Journal*, 34(20):1475–1480, May 2013. ISSN 0195-668X, 1522-9645. doi: 10.1093/eurheartj/eht024.

[71] David P. Hill, Nico Adams, Mike Bada, Colin Batchelor, Tanya Z. Berardini, Heiko Dietze, Harold J. Drabkin, Marcus Ennis, Rebecca E. Foulger, Midori A. Harris, Janna Hastings, Namrata S. Kale, Paula de Matos, Christopher J. Mungall, Gareth Owen, Paola Roncaglia, Christoph Steinbeck, Steve Turner, and Jane Lomax. Dovetailing biology and chemistry: Integrating the Gene Ontology with the ChEBI chemical ontology. *BMC Genomics*, 14(1):513, July 2013. ISSN 1471-2164. doi: 10.1186/1471-2164-14-513.

[72] Robert Hoehndorf, Paul N. Schofield, and Georgios V. Gkoutos. The role of ontologies in biological and biomedical research: A functional perspective. *Briefings in Bioinformatics*, 16(6):1069–1080, November 2015. ISSN 1467-5463. doi: 10.1093/bib/bbv011.

[73] Robert Hoehndorf, Luke Slater, Paul N. Schofield, and Georgios V. Gkoutos. Aber-OWL: A framework for ontology-based data access in biology. *BMC Bioinformatics*, 16(1):26, January 2015. ISSN 1471-2105. doi: 10.1186/s12859-015-0456-9.

[74] Thomas Hofweber. Logic and Ontology. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, summer 2018 edition, 2018.

[75] Matthew Horridge and Sean Bechhofer. The OWL API: A Java API for OWL

ontologies. *Semantic Web*, 2(1):11–21, January 2011. ISSN 1570-0844. doi: 10.3233/SW-2011-0025.

[76] Jingshan Huang, Ming Tan, Dejing Dou, Lei He, Christopher Townsend, and Patrick J. Hayes. Ontology for microRNA target prediction in human cancer. In *Proceedings of the First ACM International Conference on Bioinformatics and Computational Biology*, pages 472–474, 2010.

[77] Yasuyuki Iguchi, Kazumi Kimura, Kazuto Kobayashi, Junya Aoki, Yuka Terasawa, Kenichiro Sakai, Junichi Uemura, and Kensaku Shibazaki. Relation of Atrial Fibrillation to Glomerular Filtration Rate. *The American Journal of Cardiology*, 102 (8):1056–1059, October 2008. ISSN 00029149. doi: 10.1016/j.amjcard.2008.06.018.

[78] Richard Jackson, Ismail Kartoglu, Clive Stringer, Genevieve Gorrell, Angus Roberts, Xingyi Song, Honghan Wu, Asha Agrawal, Kenneth Lui, Tudor Groza, Damian Lewsley, Doug Northwood, Amos Folarin, Robert Stewart, and Richard Dobson. CogStack - experiences of deploying integrated information retrieval and extraction services in a large National Health Service Foundation Trust hospital. *BMC Medical Informatics and Decision Making*, 18(1):47, June 2018. ISSN 1472-6947. doi: 10.1186/s12911-018-0623-9.

[79] Pankaj Jaiswal, Shulamit Avraham, Katica Ilic, Elizabeth A. Kellogg, Susan McCouch, Anuradha Pujar, Leonore Reiser, Seung Y. Rhee, Martin M. Sachs, and Mary Schaeffer. Plant Ontology (PO): A controlled vocabulary of plant structures and growth stages. *Comparative and functional genomics*, 6(7-8):388–397, 2005.

[80] Alistair E. W. Johnson, Tom J. Pollard, Lu Shen, Li-wei H. Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G. Mark. MIMIC-III, a freely accessible critical care database. *Scientific Data*, 3(1):1–9, May 2016. ISSN 2052-4463. doi: 10.1038/sdata.2016.35.

[81] Clement Jonquet. NCBO Annotator: Semantic Annotation of Biomedical Data. page 3.

[82] Daniel Jurafsky and James H. Martin. *Speech and Language Processing: International Edition*. Pearson, Upper Saddle River, NJ, 2 edition edition, April 2008. ISBN 978-0-13-504196-3.

[83] Yannis Kalfoglou and Marco Schorlemmer. Ontology mapping: The state of the art. /core/journals/knowledge-engineering-review/article/ontology-mapping-the-state-of-the-art/9A273424C3873243A0DC50FD43C6AD4E, January 2003.

[84] Aditya Kalyanpur, Bijan Parsia, Evren Sirin, and James Hendler. Debugging unsatisfiable classes in OWL ontologies. *Journal of Web Semantics*, 3(4):268–293, December 2005. ISSN 1570-8268. doi: 10.1016/j.websem.2005.09.005.

[85] Aditya Kalyanpur, Bijan Parsia, Matthew Horridge, and Evren Sirin. Finding All Justifications of OWL DL Entailments. In Karl Aberer, Key-Sun Choi, Natasha Noy, Dean Allemang, Kyung-Il Lee, Lyndon Nixon, Jennifer Golbeck, Peter Mika, Diana Maynard, Riichiro Mizoguchi, Guus Schreiber, and Philippe Cudré-Mauroux, editors, *The Semantic Web*, Lecture Notes in Computer Science, pages 267–280, Berlin, Heidelberg, 2007. Springer. ISBN 978-3-540-76298-0. doi: 10.1007/978-3-540-76298-0_20.

[86] Yevgeny Kazakov, Markus Krötzsch, and František Simančík. The Incredible ELK. *Journal of Automated Reasoning*, 53(1):1–61, June 2014. ISSN 1573-0670. doi: 10.1007/s10817-013-9296-3.

[87] Sebastian Köhler, Sandra C. Doelken, Christopher J. Mungall, Sebastian Bauer, Helen V. Firth, Isabelle Bailleul-Forestier, Graeme C. M. Black, Danielle L. Brown, Michael Brudno, Jennifer Campbell, David R. FitzPatrick, Janan T. Eppig, Andrew P. Jackson, Kathleen Freson, Marta Girdea, Ingo Helbig, Jane A. Hurst, Johanna Jähn, Laird G. Jackson, Anne M. Kelly, David H. Ledbetter, Sahar Mansour,

Christa L. Martin, Celia Moss, Andrew Mumford, Willem H. Ouwehand, Soo-Mi Park, Erin Rooney Riggs, Richard H. Scott, Sanjay Sisodiya, Steven Van Vooren, Ronald J. Wapner, Andrew O. M. Wilkie, Caroline F. Wright, Anneke T. Vulto-van Silfhout, Nicole de Leeuw, Bert B. A. de Vries, Nicole L. Washingthon, Cynthia L. Smith, Monte Westerfield, Paul Schofield, Barbara J. Ruef, Georgios V. Gkoutos, Melissa Haendel, Damian Smedley, Suzanna E. Lewis, and Peter N. Robinson. The Human Phenotype Ontology project: Linking molecular biology and disease through phenotype data. *Nucleic Acids Research*, 42(Database issue):D966–D974, January 2014. ISSN 0305-1048. doi: 10.1093/nar/gkt1026.

[88] Sebastian Köhler, Sandra C Doelken, Barbara J Ruef, Sebastian Bauer, Nicole Washington, Monte Westerfield, George Gkoutos, Paul Schofield, Damian Smedley, Suzanna E Lewis, Peter N Robinson, and Christopher J Mungall. Construction and accessibility of a cross-species phenotype ontology along with gene annotations for biomedical research. *F1000Research*, 2:30, January 2014. ISSN 2046-1402. doi: 10.12688/f1000research.2-30.v2.

[89] Sebastian Köhler, Nicole A. Vasilevsky, Mark Engelstad, Erin Foster, Julie McMurry, Ségolène Aymé, Gareth Baynam, Susan M. Bello, Cornelius F. Boerkoel, Kym M. Boycott, Michael Brudno, Orion J. Buske, Patrick F. Chinnery, Valentina Cipriani, Laureen E. Connell, Hugh J. S. Dawkins, Laura E. DeMare, Andrew D. Devereau, Bert B. A. de Vries, Helen V. Firth, Kathleen Freson, Daniel Greene, Ada Hamosh, Ingo Helbig, Courtney Hum, Johanna A. Jähn, Roger James, Roland Krause, Stanley J. F. Laulederkind, Hanns Lochmüller, Gholson J. Lyon, Soichi Ogishima, Annie Olry, Willem H. Ouwehand, Nikolas Pontikos, Ana Rath, Franz Schaefer, Richard H. Scott, Michael Segal, Panagiotis I. Sergouniotis, Richard Sever, Cynthia L. Smith, Volker Straub, Rachel Thompson, Catherine Turner, Ernest Turro, Marijcke W. M. Veltman, Tom Vulliamy, Jing Yu, Julie von Ziegenweidt, Andreas Zankl, Stephan Züchner, Tomasz Zemojtel, Julius O. B. Jacobsen, Tudor

Groza, Damian Smedley, Christopher J. Mungall, Melissa Haendel, and Peter N. Robinson. The Human Phenotype Ontology in 2017. *Nucleic Acids Research*, 45 (D1):D865–D876, January 2017. ISSN 0305-1048. doi: 10.1093/nar/gkw1039.

[90] Sebastian Köhler, Leigh Carmody, Nicole Vasilevsky, Julius O. B. Jacobsen, Daniel Danis, Jean-Philippe Gourdine, Michael Gargano, Nomi L. Harris, Nicolas Matentzoglu, Julie A. McMurry, David Osumi-Sutherland, Valentina Cipriani, James P. Balhoff, Tom Conlin, Hannah Blau, Gareth Baynam, Richard Palmer, Dylan Gratian, Hugh Dawkins, Michael Segal, Anna C. Jansen, Ahmed Muaz, Willie H. Chang, Jenna Bergerson, Stanley J. F. Laulederkind, Zafer Yüksel, Sergi Beltran, Alexandra F. Freeman, Panagiotis I. Sergouniotis, Daniel Durkin, Andrea L. Storm, Marc Hanauer, Michael Brudno, Susan M. Bello, Murat Sincan, Kayli Rageth, Matthew T. Wheeler, Renske Oegema, Halima Lourghi, Maria G. Della Rocca, Rachel Thompson, Francisco Castellanos, James Priest, Charlotte Cunningham-Rundles, Ayushi Hegde, Ruth C. Lovering, Catherine Hajek, Annie Olry, Luigi Notarangelo, Morgan Similuk, Xingmin A. Zhang, David Gómez-Andrés, Hanns Lochmüller, Hélène Dollfus, Sergio Rosenzweig, Shruti Marwaha, Ana Rath, Kathleen Sullivan, Cynthia Smith, Joshua D. Milner, Dorothée Leroux, Cornelius F. Boerkoel, Amy Klion, Melody C. Carter, Tudor Groza, Damian Smedley, Melissa A. Haendel, Chris Mungall, and Peter N. Robinson. Expansion of the Human Phenotype Ontology (HPO) knowledge base and resources. *Nucleic Acids Research*, 47 (D1):D1018–D1027, January 2019. ISSN 1362-4962. doi: 10.1093/nar/gky1105.

[91] V Kotsev, N Minadakis, V Papakonstantinou, O Erling, I Fundulaki, and A Kiryakov. Benchmarking RDF Query Engines: The LDBC Semantic Publishing Benchmark. page 16.

[92] Lane Deirdre A. and Lip Gregory Y.H. Use of the CHA2DS2-VASc and HAS-BLED Scores to Aid Decision Making for Thromboprophylaxis in Nonvalvular

Atrial Fibrillation. *Circulation*, 126(7):860–865, August 2012. doi: 10.1161/CIRCULATIONAHA.111.060061.

[93] Frank Po-Yen Lin, Tudor Groza, Simon Kocbek, Erick Antezana, and Richard J Epstein. The Cancer Care Treatment Outcomes Ontology (CCTO): A computable ontology for profiling treatment outcomes of patients with solid tumors. *Journal of Clinical Oncology*, 35(15_suppl):e18137–e18137, May 2017. ISSN 0732-183X. doi: 10.1200/JCO.2017.35.15_suppl.e18137.

[94] Yu Lin. Towards a Semantic Web Application: Ontology-Driven Ortholog Clustering Analysis. 2011.

[95] Manuel Lobo, Andre Lamurias, and Francisco M. Couto. Identifying Human Phenotype Terms by Combining Machine Learning and Validation Rules. https://new.hindawi.com/journals/bmri/2017/8565739/, 2017.

[96] Edward Loper and Steven Bird. NLTK: The Natural Language Toolkit. *arXiv:cs/0205028*, May 2002.

[97] Maria-Angela Losi, Sandro Betocchi, Mariano Aversa, Raffaella Lombardi, Marianna Miranda, Gianluigi D'Alessandro, Alessandra Cacace, Carlo-Gabriele Tocchetti, Giovanni Barbati, and Massimo Chiariello. Determinants of atrial fibrillation development in patients with hypertrophic cardiomyopathy. *The American Journal of Cardiology*, 94(7):895–900, October 2004. ISSN 0002-9149. doi: 10.1016/j.amjcard.2004.06.024.

[98] Dalip Mahal. A Hitting Set Tree Implementation.

[99] James Malone, Tomasz Adamusiak, Ele Holloway, and Helen Parkinson. Developing an application ontology for annotation of experimental variables – Experimental Factor Ontology. *Nature Precedings*, pages 1–1, September 2009. ISSN 1756-0357. doi: 10.1038/npre.2009.3806.1.

[100] James Malone, Tomasz Adamusiak, Ele Holloway, and Helen Parkinson. Developing an application ontology for annotation of experimental variables – Experimental Factor Ontology. *Nature Precedings*, pages 1–1, September 2009. ISSN 1756-0357. doi: 10.1038/npre.2009.3806.1.

[101] J. Manimaran and T. Velmurugan. Evaluation of lexicon- and syntax-based negation detection algorithms using clinical text data. *Bio-Algorithms and Med-Systems*, 13 (4):201–213, 2017. ISSN 1895-9091. doi: 10.1515/bams-2017-0016.

[102] Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Rose Finkel, Steven Bethard, and David McClosky. The Stanford CoreNLP Natural Language Processing Toolkit. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 55–60, June 2014. doi: 10.3115/v1/P14-5010.

[103] B. J. Maron, J. M. Gardin, J. M. Flack, S. S. Gidding, T. T. Kurosaki, and D. E. Bild. Prevalence of hypertrophic cardiomyopathy in a general population of young adults. Echocardiographic analysis of 4111 subjects in the CARDIA Study. Coronary Artery Risk Development in (Young) Adults. *Circulation*, 92(4):785–789, August 1995. ISSN 0009-7322. doi: 10.1161/01.cir.92.4.785.

[104] B. J. Maron, I. Olivotto, P. Spirito, S. A. Casey, P. Bellone, T. E. Gohman, K. J. Graham, D. A. Burton, and F. Cecchi. Epidemiology of hypertrophic cardiomyopathy-related death: Revisited in a large non-referral-based patient population. *Circulation*, 102(8):858–864, August 2000. ISSN 1524-4539. doi: 10.1161/01.cir.102.8.858.

[105] Barry J. Maron. Sudden Death in Young Competitive Athletes: Clinical, Demographic, and Pathological Profiles. *JAMA*, 276(3):199, July 1996. ISSN 0098-7484. doi: 10.1001/jama.1996.03540030033028.

[106] Barry J Maron and Martin S Maron. Hypertrophic cardiomyopathy. *The Lancet*,

381(9862):242–255, January 2013. ISSN 01406736. doi: 10.1016/S0140-6736(12)60397-3.

[107] Barry J. Maron, William J. McKenna, Gordon K. Danielson, Lukas J. Kappenberger, Horst J. Kuhn, Christine E. Seidman, Pravin M. Shah, William H. Spencer, Paolo Spirito, Folkert J. Ten Cate, E. Douglas Wigle, Robert A. Vogel, Jonathan Abrams, Eric R. Bates, Bruce R. Brodie, Peter G. Danias, Gabriel Gregoratos, Mark A. Hlatky, Judith S. Hochman, Sanjiv Kaul, Robert C. Lichtenberg, Jonathan R. Lindner, Robert A. O'rourke, Gerald M. Pohost, Richard S. Schofield, Cynthia M. Tracy, William L. Winters, Werner W. Klein, Silvia G. Priori, Angeles Alonso-Garcia, Carina Blomström-Lundqvist, Guy De Backer, Jaap Deckers, Markus Flather, Jaromir Hradec, Ali Oto, Alexander Parkhomenko, Sigmund Silber, and Adam Torbicki. American College of Cardiology/European Society of Cardiology Clinical Expert Consensus Document on Hypertrophic Cardiomyopathy: A report of the American College of Cardiology Foundation Task Force on Clinical Expert Consensus Documents and the European Society of Cardiology Committee for Practice Guidelines. *Journal of the American College of Cardiology*, 42(9): 1687–1713, November 2003. ISSN 0735-1097. doi: 10.1016/S0735-1097(03)00941-0.

[108] Martin S. Maron, Ethan J. Rowin, Benjamin S. Wessler, Paula J. Mooney, Amber Fatima, Parth Patel, Benjamin C. Koethe, Mikhail Romashko, Mark S. Link, and Barry J. Maron. Enhanced American College of Cardiology/American Heart Association Strategy for Prevention of Sudden Cardiac Death in High-Risk Patients With Hypertrophic Cardiomyopathy. *JAMA Cardiology*, 4(7):644–657, July 2019. ISSN 2380-6583. doi: 10.1001/jamacardio.2019.1391.

[109] Robert T. Means. Free and Easy? Red Cell Distribution Width (RDW) and Prognosis in Cardiac Disease. *Journal of Cardiac Failure*, 17(4):299–300, April 2011. ISSN 1071-9164, 1532-8414. doi: 10.1016/j.cardfail.2011.01.008.

[110] Saeed Mehrabi, Anand Krishnan, Sunghwan Sohn, Alexandra M. Roch, Heidi

Schmidt, Joe Kesterson, Chris Beesley, Paul Dexter, C. Max Schmidt, Hongfang Liu, and Mathew Palakal. DEEPEN: A negation detection system for clinical text incorporating dependency relation into NegEx. *Journal of Biomedical Informatics*, 54:213–219, April 2015. ISSN 1532-0480. doi: 10.1016/j.jbi.2015.02.010.

[111] Paola Melacini, Cristina Basso, Annalisa Angelini, Chiara Calore, Fabiana Bobbo, Barbara Tokajuk, Nicoletta Bellini, Gessica Smaniotto, Mauro Zucchetto, Sabino Iliceto, Gaetano Thiene, and Barry J. Maron. Clinicopathological profiles of progressive heart failure in hypertrophic cardiomyopathy. *European Heart Journal*, 31 (17):2111–2123, September 2010. ISSN 0195-668X. doi: 10.1093/eurheartj/ehq136.

[112] Eric Miller. An Introduction to the Resource Description Framework. *Bulletin of the American Society for Information Science and Technology*, 25(1):15–19, 1998. ISSN 1550-8366. doi: 10.1002/bult.105.

[113] Christopher J. Mungall, Carlo Torniai, Georgios V. Gkoutos, Suzanna E. Lewis, and Melissa A. Haendel. Uberon, an integrative multi-species anatomy ontology. *Genome Biology*, 13(1):R5, January 2012. ISSN 1474-760X. doi: 10.1186/gb-2012-13-1-r5.

[114] David Nadeau and Satoshi Sekine. A survey of named entity recognition and classification. *Lingvisticæ Investigationes*, 30(1):3–26, January 2007. ISSN 0378-4169, 1569-9927. doi: 10.1075/li.30.1.03nad.

[115] Aurélie Névéol, Julien Grosjean, Stéfan Darmoni, and Pierre Zweigenbaum. Language Resources for French in the Biomedical Domain. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 2146–2151, Reykjavik, Iceland, May 2014. European Language Resources Association (ELRA).

[116] Yuji Nishizaki and Hiroyuki Daida. Red Blood Cell Distribution Width for Heart

Failure. *Internal Medicine*, 52(3):417–417, 2013. doi: 10.2169/internalmedicine.52.9214.

[117] Natalya F Noy, Monica Crubezy, Ray W Fergerson, Holger Knublauch, Samson W Tu, Jennifer Vendetti, and Mark A Musen. Protégé-2000: An Open-Source Ontology-Development and Knowledge-Acquisition Environment. page 2.

[118] Natalya F. Noy, Monica Crubézy, Ray W. Fergerson, Holger Knublauch, Samson W. Tu, Jennifer Vendetti, and Mark A. Musen. Protégé-2000: An Open-Source Ontology-Development and Knowledge-Acquisition Environment. *AMIA Annual Symposium Proceedings*, 2003:953, 2003. ISSN 1942-597X.

[119] I. Olivotto, F. Cecchi, S. A. Casey, A. Dolara, J. H. Traverse, and B. J. Maron. Impact of atrial fibrillation on the clinical course of hypertrophic cardiomyopathy. *Circulation*, 104(21):2517–2524, November 2001. ISSN 1524-4539. doi: 10.1161/hc4601.097997.

[120] C. O'Mahony, F. Jichi, M. Pavlou, L. Monserrat, A. Anastasakis, C. Rapezzi, E. Biagini, J. R. Gimeno, G. Limongelli, W. J. McKenna, R. Z. Omar, P. M. Elliott, and for the Hypertrophic Cardiomyopathy Outcomes Investigators. A novel clinical risk prediction model for sudden cardiac death in hypertrophic cardiomyopathy (HCM Risk-SCD). *European Heart Journal*, 35(30):2010–2020, August 2014. ISSN 0195-668X, 1522-9645. doi: 10.1093/eurheartj/eht439.

[121] World Health Organization, editor. *International Statistical Classification of Diseases and Related Health Problems*. World Health Organization, Geneva, 10th revision, 2nd edition edition, 2004. ISBN 978-92-4-154649-2 978-92-4-154653-9 978-92-4-154654-6.

[122] Irene Papatheodorou, Nuno A. Fonseca, Maria Keays, Y. Amy Tang, Elisabet Barrera, Wojciech Bazant, Melissa Burke, Anja Füllgrabe, Alfonso Muñoz-Pomer Fuentes, Nancy George, Laura Huerta, Satu Koskinen, Suhaib Mohammed,

Matthew Geniza, Justin Preece, Pankaj Jaiswal, Andrew F. Jarnuczak, Wolfgang Huber, Oliver Stegle, Juan Antonio Vizcaino, Alvis Brazma, and Robert Petryszak. Expression Atlas: Gene and protein expression across multiple studies and organisms. *Nucleic Acids Research*, 46(D1):D246–D251, January 2018. ISSN 0305-1048. doi: 10.1093/nar/gkx1158.

[123] Joseph E. Parrillo. Heart Disease and the Eosinophil. *New England Journal of Medicine*, 323(22):1560–1561, November 1990. ISSN 0028-4793. doi: 10.1056/NEJM199011293232211.

[124] Peter Peduzzi, John Concato, Elizabeth Kemper, Theodore R. Holford, and Alvan R. Feinstein. A simulation study of the number of events per variable in logistic regression analysis. *Journal of Clinical Epidemiology*, 49(12):1373–1379, December 1996. ISSN 0895-4356, 1878-5921. doi: 10.1016/S0895-4356(96)00236-3.

[125] Yifan Peng, Xiaosong Wang, Le Lu, Mohammadhadi Bagheri, Ronald Summers, and Zhiyong Lu. NegBio: A high-performance tool for negation and uncertainty detection in radiology reports. *AMIA Summits on Translational Science Proceedings*, 2018:188–196, May 2018. ISSN 2153-4063.

[126] Catia Pesquita, Daniel Faria, Cosmin Stroe, Emanuel Santos, Isabel F. Cruz, and Francisco M. Couto. What's in a 'nym'? Synonyms in Biomedical Ontology Matching. In Harith Alani, Lalana Kagal, Achille Fokoue, Paul Groth, Chris Biemann, Josiane Xavier Parreira, Lora Aroyo, Natasha Noy, Chris Welty, and Krzysztof Janowicz, editors, *The Semantic Web – ISWC 2013*, Lecture Notes in Computer Science, pages 526–541, Berlin, Heidelberg, 2013. Springer. ISBN 978-3-642-41335-3. doi: 10.1007/978-3-642-41335-3_33.

[127] Brock Polnaszek, Andrea Gilmore-Bykovskyi, Melissa Hovanes, Rachel Roiland, Patrick Ferguson, Roger Brown, and Amy JH Kind. Overcoming the Challenges of Unstructured Data in Multi-site, Electronic Medical Record-based Abstraction.

Medical care, 54(10):e65–e72, October 2016. ISSN 0025-7079. doi: 10.1097/MLR.0000000000000108.

[128] Luis Ramos, Richard Gil, Dimitra Anastasiou, and Maria J. Martin-Bautista. Towards a Machine of a Process (MOP) ontology to facilitate e-commerce of industrial machinery. *Computers in industry*, 65(1):108–115, 2014.

[129] reality. Reality/synonym_expansion_validation: True thesis version. Zenodo, August 2020.

[130] Raymond Reiter. A theory of diagnosis from first principles. *Artificial Intelligence*, 32(1):57–95, April 1987. ISSN 0004-3702. doi: 10.1016/0004-3702(87)90062-2.

[131] Philippe Rocca-Serra, Alvis Brazma, Helen Parkinson, Ugis Sarkans, Mohammadreza Shojatalab, Sergio Contrino, Jaak Vilo, Niran Abeygunawardena, Gaurab Mukherjee, Ele Holloway, Misha Kapushesky, Patrick Kemmeren, Gonzalo Garcia Lara, Ahmet Oezcimen, and Susanna-Assunta Sansone. ArrayExpress: A public database of gene expression data at EBI. *Comptes Rendus Biologies*, 326(10):1075–1078, October 2003. ISSN 1631-0691. doi: 10.1016/j.crvi.2003.09.026.

[132] Punnaivanam Sankar and Gnanasekaran Aghila. Design and development of chemical ontologies for reaction representation. *Journal of chemical information and modeling*, 46(6):2355–2368, 2006.

[133] Savas Sarıkaya, Şafak Şahin, Lütfi Akyol, Elif Börekçi, Yunus Keser Yılmaz, Fatih Altunkaş, and Kayıhan Karaman. Is there any relationship between RDW levels and atrial fibrillation in hypertensive patient? *African Health Sciences*, 14(1):267, March 2014. ISSN 1680-6905. doi: 10.4314/ahs.v14i1.41.

[134] Sirarat Sarntivijai, Drashtti Vasant, Simon Jupp, Gary Saunders, A. Patrícia Bento, Daniel Gonzalez, Joanna Betts, Samiul Hasan, Gautier Koscielny, Ian Dunham, Helen Parkinson, and James Malone. Linking rare and common disease:

Mapping clinical disease-phenotypes to ontologies in therapeutic target validation. *Journal of Biomedical Semantics*, 7(1):8, March 2016. ISSN 2041-1480. doi: 10.1186/s13326-016-0051-7.

[135] Slava Sazonau, Uli Sattler, and Gavin Brown. Predicting Performance of OWL Reasoners: Locally or Globally? July 2014.

[136] Lynn Marie Schriml, Cesar Arze, Suvarna Nadendla, Yu-Wei Wayne Chang, Mark Mazaitis, Victor Felix, Gang Feng, and Warren Alden Kibbe. Disease Ontology: A backbone for disease semantic integration. *Nucleic Acids Research*, 40(D1):D940–D946, January 2012. ISSN 0305-1048. doi: 10.1093/nar/gkr972.

[137] Erik Segerdell, Jeff B. Bowes, Nicolas Pollet, and Peter D. Vize. An ontology for Xenopus anatomy and development. *BMC developmental biology*, 8(1):92, 2008.

[138] Kent A. Shefchek, Nomi L. Harris, Michael Gargano, Nicolas Matentzoglu, Deepak Unni, Matthew Brush, Daniel Keith, Tom Conlin, Nicole Vasilevsky, Xingmin Aaron Zhang, James P. Balhoff, Larry Babb, Susan M. Bello, Hannah Blau, Yvonne Bradford, Seth Carbon, Leigh Carmody, Lauren E. Chan, Valentina Cipriani, Alayne Cuzick, Maria D. Rocca, Nathan Dunn, Shahim Essaid, Petra Fey, Chris Grove, Jean-Phillipe Gourdine, Ada Hamosh, Midori Harris, Ingo Helbig, Maureen Hoatlin, Marcin Joachimiak, Simon Jupp, Kenneth B. Lett, Suzanna E. Lewis, Craig McNamara, Zoë M. Pendlington, Clare Pilgrim, Tim Putman, Vida Ravanmehr, Justin Reese, Erin Riggs, Sofia Robb, Paola Roncaglia, James Seager, Erik Segerdell, Morgan Similuk, Andrea L. Storm, Courtney Thaxon, Anne Thessen, Julius O. B. Jacobsen, Julie A. McMurry, Tudor Groza, Sebastian Köhler, Damian Smedley, Peter N. Robinson, Christopher J. Mungall, Melissa A. Haendel, Monica C. Munoz-Torres, and David Osumi-Sutherland. The Monarch Initiative in 2019: An integrative data and analytic platform connecting phenotypes to genotypes across species. *Nucleic Acids Research*, 48(D1):D704–D715, January 2020. ISSN 0305-1048. doi: 10.1093/nar/gkz997.

[139] Amit G. Singal, Ashin Mukherjee, B. Joseph Elmunzer, Peter DR Higgins, Anna S. Lok, Ji Zhu, Jorge A Marrero, and Akbar K Waljee. Machine Learning Algorithms Outperform Conventional Regression Models in Predicting Development of Hepatocellular Carcinoma. *The American journal of gastroenterology*, 108(11):1723–1730, November 2013. ISSN 0002-9270. doi: 10.1038/ajg.2013.332.

[140] Siontis Konstantinos C., Geske Jeffrey B., Ong Kevin, Nishimura Rick A., Ommen Steve R., and Gersh Bernard J. Atrial Fibrillation in Hypertrophic Cardiomyopathy: Prevalence, Clinical Correlations, and Mortality in a Large High-Risk Population. *Journal of the American Heart Association*, 3(3):e001002. doi: 10.1161/JAHA.114.001002.

[141] Nicholas Sioutos, Sherri de Coronado, Margaret W. Haber, Frank W. Hartel, Wen-Ling Shaiu, and Lawrence W. Wright. NCI Thesaurus: A semantic model integrating cancer-related clinical and molecular information. *Journal of Biomedical Informatics*, 40(1):30–43, February 2007. ISSN 1532-0464. doi: 10.1016/j.jbi.2006.02.013.

[142] Luke Slater, Georgios V. Gkoutos, Paul N. Schofield, and Robert Hoehndorf. To MIREOT or not to MIREOT? A Case Study of the Impact of Using MIREOT in the Experimental Factor Ontology (EFO). In *ICBO/BioCreative*, 2016.

[143] Luke T. Slater, William Bradlow, Robert Hoehndorf, Dino FA Motti, Simon Ball, and Georgios V. Gkoutos. Komenti: A semantic text mining framework. *bioRxiv*, page 2020.08.04.233049, August 2020. doi: 10.1101/2020.08.04.233049.

[144] Barry Smith and Chris Welty. Ontology: Towards a new synthesis. 2001.

[145] Barry Smith, Anand Kumar, and Thomas Bittner. Basic Formal Ontology for bioinformatics. https://philarchive.org, 2005.

[146] Cynthia L. Smith, Carroll-Ann W. Goldsmith, and Janan T. Eppig. The Mammalian Phenotype Ontology as a tool for annotating, analyzing and comparing phe-

notypic information. *Genome Biology*, 6(1):R7, December 2004. ISSN 1474-760X. doi: 10.1186/gb-2004-6-1-r7.

[147] Sunghwan Sohn, Stephen Wu, and Christopher G. Chute. Dependency Parser-based Negation Detection in Clinical Narratives. *AMIA Summits on Translational Science Proceedings*, 2012:1–8, March 2012. ISSN 2153-4063.

[148] Irena Spasic, Sophia Ananiadou, John McNaught, and Anand Kumar. Text mining and ontologies in biomedicine: Making sense of raw text. *Briefings in Bioinformatics*, 6(3):239–251, September 2005. ISSN 1467-5463. doi: 10.1093/bib/6.3.239.

[149] Judy Sprague, Leyla Bayraktaroglu, Dave Clements, Tom Conlin, David Fashena, Ken Frazer, Melissa Haendel, Douglas G. Howe, Prita Mani, and Sridhar Ramachandran. The Zebrafish Information Network: The zebrafish model organism database. *Nucleic acids research*, 34(suppl_1):D581–D585, 2006.

[150] Ewout Steyerberg. *Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating*. Statistics for Biology and Health. Springer International Publishing, second edition, 2019. ISBN 978-3-030-16398-3. doi: 10.1007/978-3-030-16399-0.

[151] Ewout W. Steyerberg, Andrew J. Vickers, Nancy R. Cook, Thomas Gerds, Mithat Gonen, Nancy Obuchowski, Michael J. Pencina, and Michael W. Kattan. Assessing the performance of prediction models: A framework for some traditional and novel measures. *Epidemiology (Cambridge, Mass.)*, 21(1):128–138, January 2010. ISSN 1044-3983. doi: 10.1097/EDE.0b013e3181c30fb2.

[152] Stroke Risk in Atrial Fibrillation Working Group. Independent predictors of stroke in patients with atrial fibrillation: A systematic review. *Neurology*, 69(6):546–554, August 2007. ISSN 1526-632X. doi: 10.1212/01.wnl.0000267275.68538.8d.

[153] Wencheng Sun, Zhiping Cai, Yangyang Li, Fang Liu, Shengqun Fang, and Guoyan

Wang. Data Processing and Text Mining Technologies on Electronic Medical Records: A Review. https://www.hindawi.com/journals/jhe/2018/4302425/, 2018.

[154] Maria Taboada, Hadriana Rodriguez, Ranga C. Gudivada, and Diego Martinez. A new synonym-substitution method to enrich the human phenotype ontology. *BMC Bioinformatics*, 18, October 2017. ISSN 1471-2105. doi: 10.1186/s12859-017-1858-7.

[155] Maxwell Taggart, Wendy W. Chapman, Benjamin A. Steinberg, Shane Ruckel, Arianna Pregenzer-Wenzler, Yishuai Du, Jeffrey Ferraro, Brian T. Bucher, Donald M. Lloyd-Jones, Matthew T. Rondina, and Rashmee U. Shah. Comparison of 2 Natural Language Processing Methods for Identification of Bleeding Among Critically Ill Patients. *JAMA Network Open*, 1(6):e183451, October 2018. ISSN 2574-3805. doi: 10.1001/jamanetworkopen.2018.3451.

[156] Tomoko Tani, Kazuaki Tanabe, Miwa Ono, Kazuto Yamaguchi, Midori Okada, Toshiaki Sumida, Toshiko Konda, Yoko Fujii, Junichi Kawai, Toshikazu Yagi, Masatake Sato, Motoaki Ibuki, Minako Katayama, Koichi Tamita, Kenji Yamabe, Atsushi Yamamuro, Kunihiko Nagai, Kenichi Shiratori, and Shigefumi Morioka. Left atrial volume and the risk of paroxysmal atrial fibrillation in patients with hypertrophic cardiomyopathy. *Journal of the American Society of Echocardiography*, 17 (6):644–648, June 2004. ISSN 0894-7317. doi: 10.1016/j.echo.2004.02.010.

[157] Hideyuki Tanushi, Hercules Dalianis, Martin Duneld, Maria Kvist, Maria Skeppstedt, and Sumithra Velupillai. Negation Scope Delimitation in Clinical Text Using Three Approaches: NegEx, PyConTextNLP and SynNeg. In *Proceedings of the 19th Nordic Conference of Computational Linguistics (NODALIDA 2013)*, pages 387–397, Oslo, Norway, May 2013. Linköping University Electronic Press, Sweden.

[158] Stuart J. Taylor and Sanda M. Harabagiu. The Role of a Deep-Learning Method for Negation Detection in Patient Cohort Identification from Electroencephalography

Reports. *AMIA Annual Symposium Proceedings*, 2018:1018–1027, December 2018. ISSN 1942-597X.

[159] Donald Teare. ASYMMETRICAL HYPERTROPHY OF THE HEART IN YOUNG ADULTS. *British Heart Journal*, 20(1):1–8, January 1958. ISSN 0007-0769.

[160] The OBI Consortium, Barry Smith, Michael Ashburner, Cornelius Rosse, Jonathan Bard, William Bug, Werner Ceusters, Louis J Goldberg, Karen Eilbeck, Amelia Ireland, Christopher J Mungall, Neocles Leontis, Philippe Rocca-Serra, Alan Ruttenberg, Susanna-Assunta Sansone, Richard H Scheuermann, Nigam Shah, Patricia L Whetzel, and Suzanna Lewis. The OBO Foundry: Coordinated evolution of ontologies to support biomedical data integration. *Nature Biotechnology*, 25(11): 1251–1255, November 2007. ISSN 1087-0156, 1546-1696. doi: 10.1038/nbt1346.

[161] Terry Therneau, Cynthia Crowson, and Elizabeth Atkinson. Using Time Dependent Covariates and Time Dependent Coefficients in the Cox Model. page 27.

[162] Syed Hamid Tirmizi, Stuart Aitken, Dilvan A. Moreira, Chris Mungall, Juan Sequeda, Nigam H. Shah, and Daniel P. Miranker. Mapping between the OBO and OWL ontology languages. *Journal of Biomedical Semantics*, 2(1):S3, March 2011. ISSN 2041-1480. doi: 10.1186/2041-1480-2-S1-S3.

[163] Jack V. Tu. Advantages and disadvantages of using artificial neural networks versus logistic regression for predicting medical outcomes. *Journal of Clinical Epidemiology*, 49(11):1225–1231, November 1996. ISSN 0895-4356, 1878-5921. doi: 10.1016/S0895-4356(96)00002-9.

[164] Tania Tudorache, Jennifer Vendetti, and Natasha Noy. Web-Protege: A Lightweight OWL Ontology Editor for the Web. In *Fifth OWLED Workshop on OWL: Experiences and Directions*, January 2008.

[165] Ilinca Tudose, Janna Hastings, Venkatesh Muthukrishnan, Gareth Owen, Steve Turner, Adriano Dekker, Namrata Kale, Marcus Ennis, and Christoph Steinbeck.

OntoQuery: Easy-to-use web-based OWL querying. *Bioinformatics*, 29(22):2955–2957, November 2013. ISSN 1367-4803. doi: 10.1093/bioinformatics/btt514.

[166] Udelson James E. Evaluating and Reducing the Risk of Sudden Death in Hypertrophic Cardiomyopathy. *Circulation*, 139(6):727–729, February 2019. doi: 10.1161/CIRCULATIONAHA.118.038436.

[167] Tomoya Ueda, Rika Kawakami, Manabu Horii, Yu Sugawara, Takaki Matsumoto, Sadanori Okada, Taku Nishida, Tsunenari Soeda, Satoshi Okayama, Satoshi Somekawa, Yukiji Takeda, Makoto Watanabe, Hiroyuki Kawata, Shiro Uemura, and Yoshihiko Saito. High Mean Corpuscular Volume Is a New Indicator of Prognosis in Acute Decompensated Heart Failure. *Circulation Journal*, advpub, 2013. doi: 10.1253/circj.CJ-13-0718.

[168] Ramachandran S. Vasan, Emelia J. Benjamin, and Daniel Levy. Prevalence, clinical features and prognosis of diastolic heart failure: An epidemiologic perspective. *Journal of the American College of Cardiology*, 26(7):1565–1574, December 1995. ISSN 0735-1097, 1558-3597. doi: 10.1016/0735-1097(95)00381-9.

[169] Denny Vrandečić and Markus Krötzsch. Wikidata: A free collaborative knowledgebase. *Communications of the ACM*, 57(10):78–85, September 2014. ISSN 00010782. doi: 10.1145/2629489.

[170] Yiwen Wang, Natalia Stash, Lora Aroyo, Peter Gorgels, Lloyd Rutledge, and Guus Schreiber. Recommendations based on semantically enriched museum collections. *Journal of Web Semantics*, 6(4):283–290, November 2008. ISSN 15708268. doi: 10.1016/j.websem.2008.09.002.

[171] E. D. Wigle, H. Rakowski, B. P. Kimball, and W. G. Williams. Hypertrophic cardiomyopathy. Clinical spectrum and treatment. *Circulation*, 92(7):1680–1692, October 1995. ISSN 0009-7322. doi: 10.1161/01.cir.92.7.1680.

[172] Bryan Williams, Giuseppe Mancia, Wilko Spiering, Enrico Agabiti Rosei, Michel Azizi, Michel Burnier, Denis L. Clement, Antonio Coca, Giovanni de Simone, Anna Dominiczak, Thomas Kahan, Felix Mahfoud, Josep Redon, Luis Ruilope, Alberto Zanchetti, Mary Kerins, Sverre E. Kjeldsen, Reinhold Kreutz, Stephane Laurent, Gregory Y. H. Lip, Richard McManus, Krzysztof Narkiewicz, Frank Ruschitzka, Roland E. Schmieder, Evgeny Shlyakhto, Costas Tsioufis, Victor Aboyans, Ileana Desormais, Guy De Backer, Anthony M. Heagerty, Stefan Agewall, Murielle Bochud, Claudio Borghi, Pierre Boutouyrie, Jana Brguljan, Héctor Bueno, Enrico G. Caiani, Bo Carlberg, Neil Chapman, Renata Cífková, John G. F. Cleland, Jean-Philippe Collet, Ioan Mircea Coman, Peter W. de Leeuw, Victoria Delgado, Paul Dendale, Hans-Christoph Diener, Maria Dorobantu, Robert Fagard, Csaba Farsang, Marc Ferrini, Ian M. Graham, Guido Grassi, Hermann Haller, F. D. Richard Hobbs, Bojan Jelakovic, Catriona Jennings, Hugo A. Katus, Abraham A. Kroon, Christophe Leclercq, Dragan Lovic, Empar Lurbe, Athanasios J. Manolis, Theresa A. McDonagh, Franz Messerli, Maria Lorenza Muiesan, Uwe Nixdorff, Michael Hecht Olsen, Gianfranco Parati, Joep Perk, Massimo Francesco Piepoli, Jorge Polonia, Piotr Ponikowski, Dimitrios J. Richter, Stefano F. Rimoldi, Marco Roffi, Naveed Sattar, Petar M. Seferovic, Iain A. Simpson, Miguel Sousa-Uva, Alice V. Stanton, Philippe van de Borne, Panos Vardas, Massimo Volpe, Sven Wassmann, Stephan Windecker, Jose Luis Zamorano, Stephan Windecker, Victor Aboyans, Stefan Agewall, Emanuele Barbato, Héctor Bueno, Antonio Coca, Jean-Philippe Collet, Ioan Mircea Coman, Veronica Dean, Victoria Delgado, Donna Fitzsimons, Oliver Gaemperli, Gerhard Hindricks, Bernard Iung, Peter Jüni, Hugo A. Katus, Juhani Knuuti, Patrizio Lancellotti, Christophe Leclercq, Theresa A. McDonagh, Massimo Francesco Piepoli, Piotr Ponikowski, Dimitrios J. Richter, Marco Roffi, Evgeny Shlyakhto, Iain A. Simpson, Miguel Sousa-Uva, Jose Luis Zamorano, Costas Tsioufis, Empar Lurbe, Reinhold Kreutz, Murielle Bochud, Enrico Agabiti Rosei, Bojan Jelakovic, Michel Azizi, Andrzej Januszewics, Thomas Kahan, Jorge Polonia,

Philippe van de Borne, Bryan Williams, Claudio Borghi, Giuseppe Mancia, Gianfranco Parati, Denis L. Clement, Antonio Coca, Athanasios Manolis, Dragan Lovic, Salim Benkhedda, Parounak Zelveian, Peter Siostrzonek, Ruslan Najafov, Olga Pavlova, Michel De Pauw, Larisa Dizdarevic-Hudic, Dimitar Raev, Nikos Karpettas, Aleš Linhart, Michael Hecht Olsen, Amin Fouad Shaker, Margus Viigimaa, Kaj Metsärinne, Marija Vavlukis, Jean-Michel Halimi, Zurab Pagava, Heribert Schunkert, Costas Thomopoulos, Dénes Páll, Karl Andersen, Michael Shechter, Giuseppe Mercuro, Gani Bajraktari, Tatiana Romanova, Kārlis Trušinskis, Georges A. Saade, Gintare Sakalyte, Stéphanie Noppe, Daniela Cassar DeMarco, Alexandru Caraus, Janneke Wittekoek, Tonje Amb Aksnes, Piotr Jankowski, Jorge Polonia, Dragos Vinereanu, Elena I. Baranova, Marina Foscoli, Ana Djordjevic Dikic, Slavomira Filipova, Zlatko Fras, Vicente Bertomeu-Martínez, Bo Carlberg, Thilo Burkard, Wissem Sdiri, Sinan Aydogdu, Yuriy Sirenko, Adrian Brady, Thomas Weber, Irina Lazareva, Tine De Backer, Sekib Sokolovic, Bojan Jelakovic, Jiri Widimsky, Margus Viigimaa, Ilkka Pörsti, Thierry Denolle, Bernhard K. Krämer, George S. Stergiou, Gianfranco Parati, Kārlis Trušinskis, Marius Miglinas, Eva Gerdts, Andrzej Tykarski, Manuel de Carvalho Rodrigues, Maria Dorobantu, Irina Chazova, Dragan Lovic, Slavomira Filipova, Jana Brguljan, Julian Segura, Anders Gottsäter, Antoinette Pechère-Bertschi, Serap Erdine, Yuriy Sirenko, and Adrian Brady. 2018 ESC/ESH Guidelines for the management of arterial hypertension-The Task Force for the management of arterial hypertension of the European Society of Cardiology (ESC) and the European Society of Hypertension (ESH). *European Heart Journal*, 39(33):3021–3104, September 2018. ISSN 0195-668X. doi: 10.1093/eurheartj/ehy339.

[173] Stephen Wu, Timothy Miller, James Masanz, Matt Coarr, Scott Halgrim, David Carrell, and Cheryl Clark. Negation's Not Solved: Generalizability Versus Optimizability in Clinical Natural Language Processing. *PLoS ONE*, 9(11), November 2014. ISSN 1932-6203. doi: 10.1371/journal.pone.0112774.

[174] Hong-Jie Yang, Xin Liu, Chuan Qu, Shao-Bo Shi, Jin-Jun Liang, and Bo Yang. Usefulness of Red Blood Cell Distribution Width to Predict Heart Failure Hospitalization in Patients with Hypertrophic Cardiomyopathy. *International Heart Journal*, 59(4):779–785, 2018. doi: 10.1536/ihj.17-507.

[175] Erfan Younesi, Sam Ansari, Michaela Guendel, Shiva Ahmadi, Chris Coggins, Julia Hoeng, Martin Hofmann-Apitius, and Manuel C. Peitsch. CSEO – the Cigarette Smoke Exposure Ontology. *Journal of Biomedical Semantics*, 5(1):31, July 2014. ISSN 2041-1480. doi: 10.1186/2041-1480-5-31.

[176] Jie Zheng, Marcelline R. Harris, Anna Maria Masci, Yu Lin, Alfred Hero, Barry Smith, and Yongqun He. The Ontology of Biological and Clinical Statistics (OBCS) for standardized and reproducible statistical analysis. *Journal of Biomedical Semantics*, 7(1):53, September 2016. ISSN 2041-1480. doi: 10.1186/s13326-016-0100-2.

[177] Yang Zhou, Rong Jin, and Steven Chu-Hong Hoi. Exclusive Lasso for Multi-task Feature Selection. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, pages 988–995, March 2010.