Article published online: 2022-07-27

Evaluating the Effect of a COVID-19 Predictive Model to Facilitate Discharge: A Randomized Controlled Trial

Vincent J. Major[1] Simon A. Jones[1] Narges Razavian[1] Ashley Bagheri[1] Felicia Mendoza[1] Jay Stadelman[1] Leora I. Horwitz[1,2] Jonathan Austrian[2] Yindalon Aphinyanaphongs[1]

[1] Center for Healthcare Innovation & Delivery Science, Department of Population Health, NYU Grossman School of Medicine, New York, New York, United States
[2] Department of Medicine, NYU Grossman School of Medicine, New York, New York, United States

Address for correspondence Vincent J. Major, PhD, NYU Grossman School of Medicine, New York, NY 10016, United States (e-mail: vincent.major@nyulangone.org).

Appl Clin Inform 2022;13:632–640.

Abstract

Background We previously developed and validated a predictive model to help clinicians identify hospitalized adults with coronavirus disease 2019 (COVID-19) who may be ready for discharge given their low risk of adverse events. Whether this algorithm can prompt more timely discharge for stable patients in practice is unknown.

Objectives The aim of the study is to estimate the effect of displaying risk scores on length of stay (LOS).

Methods We integrated model output into the electronic health record (EHR) at four hospitals in one health system by displaying a green/orange/red score indicating low/moderate/high-risk in a patient list column and a larger COVID-19 summary report visible for each patient. Display of the score was pseudo-randomized 1:1 into intervention and control arms using a patient identifier passed to the model execution code. Intervention effect was assessed by comparing LOS between intervention and control groups. Adverse safety outcomes of death, hospice, and re-presentation were tested separately and as a composite indicator. We tracked adoption and sustained use through daily counts of score displays.

Results Enrolling 1,010 patients from May 15, 2020 to December 7, 2020, the trial found no detectable difference in LOS. The intervention had no impact on safety indicators of death, hospice or re-presentation after discharge. The scores were displayed consistently throughout the study period but the study lacks a causally linked process measure of provider actions based on the score. Secondary analysis revealed complex dynamics in LOS temporally, by primary symptom, and hospital location.

Conclusion An AI-based COVID-19 risk score displayed passively to clinicians during routine care of hospitalized adults with COVID-19 was safe but had no detectable impact on LOS. Health technology challenges such as insufficient adoption, nonuniform use, and provider trust compounded with temporal factors of the COVID-19 pandemic may have contributed to the null result.

Trial registration ClinicalTrials.gov identifier: NCT04570488.

Keywords

- COVID-19
- randomized controlled trial
- pragmatic trial
- artificial intelligence
- clinical decision support

received
December 7, 2021
accepted
April 17, 2022

DOI https://doi.org/10.1055/s-0042-1750416.
ISSN 1869-0327.

Georg Thieme Verlag KG, Rüdigerstraße 14, 70469 Stuttgart, Germany

Background and Significance

As coronavirus disease 2019 (COVID-19) spread across the world, early work highlighted its differences to other respiratory diseases and implicated various prognostic factors.[1–4] Dozens and then hundreds of prognostic tools and risk scores were developed using clinical data, but few were subsequently validated (externally or prospectively) and fewer still found utility in clinical practice.[5–12] The pandemic spread unevenly across the globe with regions experiencing secondary waves driven by variants. With each passing week, care improved over time as anecdote evolved into evidence. This rapid, nonlinear, and nonuniform evolution of community pressure, variants, and care poses an unprecedented challenge to adapt—for clinicians, health care systems, policy, and most of all, artificial intelligence (AI).

A predictive model that uses recent vital signs, lab results, and current oxygen support variables was previously validated prospectively at our institution[13] and made available for others. This system was implemented in our electronic health record (EHR) and integrated into clinical care, following guidelines for operationalizing AI systems.[14,15] The system can precisely identify COVID-19 patients who are very likely to have no COVID-related adverse events in the subsequent 96 hours and, thus, may be appropriate for de-escalation of care or discharge. While previously determined to be accurate (93% positive predictive value at 68% sensitivity and 94% specificity with green alert rate of 41%), the system's ability to augment clinical care was unknown.

Consistent with the clinical expectation for rigorous evaluations of any intervention that may affect patient safety, a randomized controlled trial was designed before the system launched. In this study, we integrated scores into the EHR (Epic Systems, Verona, Wisconsin, United States), promoted the use of the score (by presenting to key stakeholder groups, broadcasting emails, and sharing tip sheets), and investigated the system's impact to shorten length of stay (LOS) for patients assessed as low risk. Two best practices of program evaluation were employed: (1) use of random allocation of intervention and control groups to mitigate temporal changes and unforeseen confounding, and (2) tracking over time of a process measure of adoption and sustained use (anonymized display metrics). While both have limitations, evaluations with these approaches are rare in the AI field despite broad recommendation from guidelines for clinical trials, implementation studies, and evaluation of health technology.[16–19]

Our hypothesis was that clinical decision support (CDS) based on the predictive score would give clinicians confidence in safely discharging patients earlier, causing a decrease in LOS and enabling the health system to better accommodate surges of patients presenting with COVID-19.

Methods

Study Design, Setting, and Subjects

The randomized controlled trial met the criteria of quality improvement by the NYU Grossman School of Medicine IRB and did not require IRB approval and participant consent was not required. The trial was registered (ClinicalTrials.gov identifier: NCT04570488) and the protocol is included in ►Supplementary Appendix A (available in the online version). Presentation of this study followed the Consolidated Standards of Reporting Trials extension for interventions involving AI (CONSORT-AI) reporting guidelines.

Automatic screening considered adults aged 18 years or older who were hospitalized between May 15 and December 7, 2020 at any of four inpatient locations of NYU Langone Health. Eligible patients had to have COVID-19 listed as an active infection and have sufficient data in their chart (vitals and lab results) to produce at least one valid score by the AI system during their admission. Patients who received ICU-level care, who died during their hospitalization, or who never had a low risk (green colored) score generated were excluded from the primary analysis (but were included when analyzing safety outcomes).

Randomization to Treatment Group

The integrated AI system included a module to randomize patients based on a patient identifier internal to our EHR and not easily visible to clinicians. This internal identifier is assigned to patients sequentially when they are first registered in the EHR. We randomly assigned patients in a 1:1 ratio to have scores displayed or marked "Hidden." Clinicians were therefore not blinded to treatment allocation, but were blinded to the risk score of the control patients. Scores and colors were generated for the control group in the background but never shown to any user.

The intervention consisted of passive CDS communicating risk scores updated every 30 minutes. Scores ranged from 0 to 100 where smaller scores indicated lower risk of adverse events and scores were binned into three color ranges: green, orange, and red, consistent with the convention of other risk scores, e.g., green for low-risk for a bad outcome. Score cutoffs were determined by an interdisciplinary group of clinicians and data scientists based on a positive predictive value of no adverse events of 90% (green) and 80% (orange and green) with expected cumulative incidence of no adverse events within these color groups as: green = 90%, orange = 67%, and red = 8%.[13]

Scores were presented to clinicians during routine delivery of care via two channels: (1) as a column that could be added to a provider's patient list, and (2) as one element of a larger COVID-19 summary report specific to each patient.[13] A user could manually add the column to a new or existing patient list for their own use or to a list shared within a team. The patient list column consists of a narrow display of a colored oval containing the numeric score that expands into an explanatory bubble when the user hovers their cursor over it. The expanded bubble contains the same information about the patient's risk score as the larger display used in the summary report: the patient's current color-coded score; a trendline of the patient's recent scores; and a table of predictive factors, their current value and percent contribution to the overall risk. Only the nine largest explanatory factors in magnitude were presented to communicate both

risk factors and protective factors. Patients allocated to the control group had a bubble with a grey "NA" score and empty components (no trend line, no table of predictive factors). A detailed description of the CDS is presented in ►Supplementary Appendix B (available in the online version).

Outreach efforts consisted of presentations at key stakeholder meetings (e.g., hospitalist meetings and service line meetings), broadcast emails, and tip sheets. The message highlighted: (1) how the model was designed and tested to support its clinical validity; (2) where the scores were visible to clinicians; and (3) recommendations on how to interpret and apply the risk scores into the clinical plan, with a focus on green-scored patients at low risk of adverse events who may be appropriate for de-escalation of care including discharge.

Outcomes and Follow-Up

Adult patients with active COVID-19 infection admitted between May 15, 2020 and December 7, 2020 across four hospitals of one academic health system were scored and screened. The primary cohort included those discharged alive with at least one green score. The primary and secondary outcomes are reduction in LOS after first green score (gLOS) and complete LOS.

Adverse safety outcomes including in-hospital death or hospice, 30-day post-discharge death or hospice, and 30-day re-presentation were monitored for the primary cohort and a larger group of all screened and scored patients (including ineligible patients who never reached a green score, who received ICU-level care, or who died). If the intervention of de-escalating care for green-scored patients led to unintended consequences such as a reduction in monitoring or premature discharge, patients may have encountered these outcomes at higher frequency.

Statistical Analysis

The primary intention-to-treat analysis estimated the median difference of gLOS using a Mann-Whitney U test (Wilcoxon rank-sum). Secondary analyses included comparing LOS with a Mann-Whitney test and safety outcomes with χ^2 tests separately and combined into a composite. A planned alternative analysis evaluated for differences in gLOS by employing a Gamma regression analysis controlling for sex, age (second order polynomial), location (one of four hospitals), primary symptoms consistent with COVID-19, and month of first green score (►Supplementary Appendix C, available in the online version).

The planned sample size was 1,000 patients (500 in each group) based on a statistical power of 90% to detect a 0.5-day median improvement in gLOS with a two-sided Mann-Whitney test ($\alpha = 0.05$). A distribution of gLOS from a retrospective cohort of patients with COVID-19 admitted between March and April, 2020 ($n = 330$) was used for simulated power calculations.

Patients in the primary cohort (i.e., scored green and discharged alive without any ICU level care) definitionally had no missing values for gLOS, LOS, or in-hospital death. The remaining 30-day safety outcomes are assumed to have not occurred in cases where no known re-presentation or death was noted given ample follow-up for all patients (minimum follow-up between discharge and data retrieval was 88 days, median [IQR] of 248 [179, 294]).

All estimates were based on two-sided tests. Statistical significance was defined as $p < 0.05$. All analyses were performed using R. version 3.5.2.[20]

Measuring Adoption and Sustained Use

Adoption and sustained use of the tool was assessed by how frequently scores were displayed via the two channels (patient list column and summary report). These data are available as anonymized metrics by our EHR vendor (Epic Systems) as daily counts of display—when the system prepares the score to be available for display—not necessarily seen by a user who may open a large list but not scroll to view all patients. Metrics were extracted in their most granular form of daily counts—where linking to the patient/provider and stratification for hospital/department were impossible. Deduplication was also not possible, one user opening a list containing two patients six times in one day contributes twelve to that day's total.

Results

Screening and Enrollment

During the study period of 207 days, 1,415 patient admissions were scored by the system (►Fig. 1) and randomized into intervention ($n = 712$) and control ($n = 703$). After omitting 405 ineligible patients, the primary cohort included a total of 1,010 patients with 513 allocated to the intervention and 497 to the control. All patients received the intervention or control as allocated.

Among enrolled patients, 526 (52%) were female, 484 (48%) were male, and the mean (SD) age was 58.9 (20.1) years (►Table 1). Other demographic and geographic characteristics were similar between the allocated groups.

Primary Outcome

Both control and intervention groups were scored every 30 minutes and had a similar LOS prior to each patient's first green score (median [IQR], control: 0.39 [0.09–1.1] vs. intervention: 0.45 [0.10–1.2] days, $p = 0.6$). The gLOS of the intervention group where scores were visible to clinicians was similar to the control group with no visible scores (►Table 2).

Secondary Outcomes

The secondary outcomes related to LOS and adverse safety outcomes were also similar between groups (►Table 2). Adverse safety outcomes were largely re-presentations to the hospital or emergency department within 30 days, followed by inpatient mortality or initiation of hospice with very few cases of post-discharge 30-day mortality.

Measuring Adoption and Sustained Use

Clinicians rapidly adopted the score into their workflow as seen by the jump in display metrics in the week including and

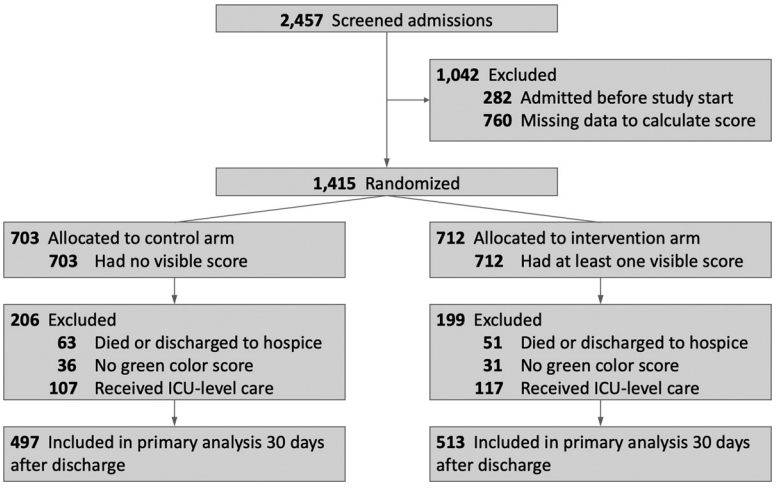

Fig. 1 CONSORT diagram. CONSORT, Consolidated Standards of Reporting Trials extension for interventions involving AI.

following launch (►**Fig. 2** colored dots with vertical line indicating launch). Over the study period, the number of admitted patients with active COVID-19 dropped and rose again in a *U*-shape (grey dots). Despite this, clinicians consistently kept the risk score column in their patient list (►**Fig. 2A**), though whether they were seen and acted upon is unclear (a passive effort by the clinical team). Displays of the patient-specific report (an active effort by the clinical team) closely follow the number of hospitalized patients with COVID-19 (►**Fig. 2B**) while preserving a ratio of approximately four displays per patient each week. Together, these metrics indicate a relatively uniform, but low, exposure likelihood—that users would have observed a patient's green score—throughout the study period. Both metrics are unscaled and give no indication of the proportion of adopters among all clinicians.

Surveying users from all departments was infeasible but adoption by one key user group was investigated in early

Table 1 Baseline characteristics of enrolled patients

	No. (%)					
	All patients (n = 1,010)		Control group (n = 497)		Intervention group (n = 513)	
Age, years						
Mean (SD)	58.9	(20.1)	58.4	(20.2)	59.4	(20.0)
Sex						
Male	484	(47.9)	223	(44.9)	261	(50.9)
Female	526	(52.1)	274	(55.1)	252	(49.1)
Ethnicity						
Hispanic	258	(25.5)	128	(25.8)	130	(25.3)
Race						
White	511	(50.6)	245	(49.3)	266	(51.9)
African American (Black)	122	(12.1)	72	(14.5)	50	(9.7)
Asian	61	(6.0)	31	(6.2)	30	(5.8)
Native American	12	(1.2)	5	(1.0)	7	(1.4)
Pacific Islander	9	(0.9)	6	(1.2)	3	(0.6)
Other	295	(29.2)	138	(27.8)	157	(30.6)
Location						
Tisch/Kimmel	270	(26.7)	140	(28.2)	130	(25.3)
Orthopedics	18	(1.8)	9	(1.8)	9	(1.8)
Brooklyn	416	(41.2)	210	(42.3)	206	(40.2)
Long Island	306	(30.3)	138	(27.8)	168	(32.7)

Table 2 Differences in primary and secondary outcomes at discharge and 30-day follow-up

	Control group		Intervention group		p-Value
Primary outcome	(*n* = 497)		(*n* = 513)		
gLOS, days, median [IQR]	3.23	[1.75–6.00]	3.18	[1.76–5.95]	0.8
Secondary outcome	(*n* = 497)		(*n* = 513)		
LOS, days, median [IQR]	4.50	[2.34–7.65]	4.20	[2.33–7.51]	0.8
Adverse safety outcomes in scored patients discharged alive.	(*n* = 640)		(*n* = 661)		
30-d re-presentation, no (%)	115	(18.0%)	138	(20.9%)	0.2
30-d mortality, no (%)	3	(0.47%)	8	(1.2%)	0.2
Adverse safety outcomes in all scored patients.	(*n* = 702)		(*n* = 713)		
Inpatient mortality/hospice, no (%)	63	(9.0%)	50	(7.0%)	0.2
Any adverse outcome, no (%)	178	(25.3%)	191	(26.8%)	0.6

Abbreviations: gLOS, length of stay after first green score; LOS, length of stay; IQR, interquartile range.

2021. From the set of 31 General Medicine hospitalists at the Tisch/Kimmel hospital, 22 regularly treated COVID-19 patients throughout the study period but only five included the risk score in a patient list. More may have sought out the score directly from the summary report but outreach and communication efforts likely did not reach, or convince, all users of the intervention's potential benefit.

Discussion

The intervention of displaying a color-coded risk score for patients with COVID-19 appears to be safe but neither efficacious nor effective in reducing LOS. A planned gamma regression also found no intervention effect (►**Supplementary Appendix C** [available in the online version]: age, one hospital location, and the second half of the study period were influential of gLOS).

There are numerous factors in model development, communication, and provider interpretation of the model scores where an insufficiency could have hindered the potential effect of the intervention to reduce gLOS. These fall into three major groups: first, discharge readiness: (1) The model was developed for COVID-19 and it overlooks the effects of additional acute or chronic conditions. (2) The system identifies COVID-19 patients at low-risk for adverse events which may not align with provider perceptions of being stable or ready for discharge. (3) Medical stability is not the sole driver of discharge readiness and this intervention does not address social needs.

The second group of factors relate to user perceptions and learning: (4) Adoption and sustained use may have been

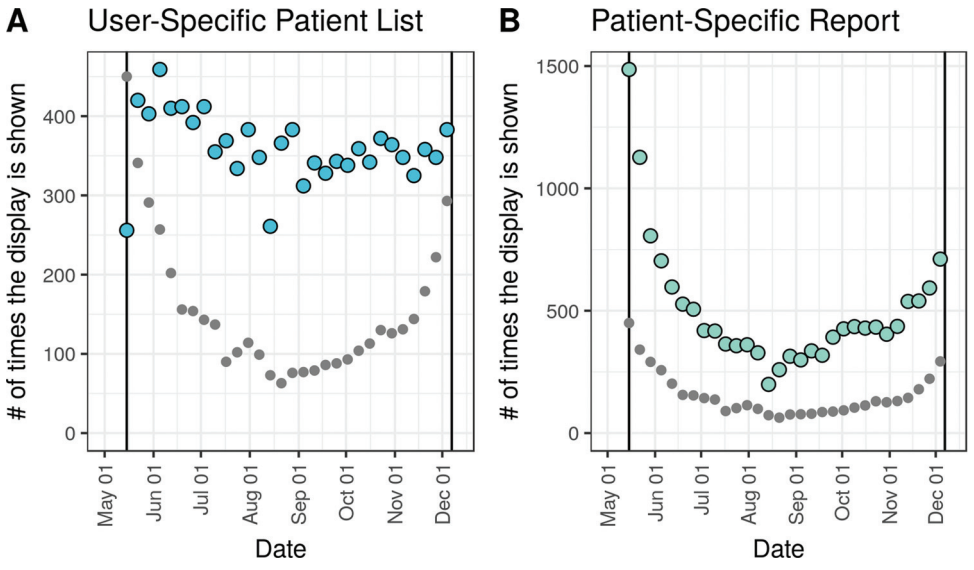

Fig. 2 Adoption and sustained use of two communication channels available to clinicians. Colored dots represent the weekly total number of times the score was displayed to any user via each display channel, either a personal or shared patient list **(A)** or the patient-specific report of COVID-19 information **(B)**. The weekly census of hospitalized COVID-19 patients is shown with grey dots for reference (calculated as the number of unique encounters scored by the system in that week).

dampened due to issues of distrust or algorithm aversion.[21] (5) A clinician's experiences caring for patients with COVID-19 involves continued learning where observing risk scores may have expedited development of a mental model of discharge readiness that would diminish the model's utility.

The third group relates to pragmatic choices of outreach and CDS integration: (6) Uptake of the intervention may have been limited by insufficient outreach in the setting of significant competing COVID-19 communications. (7) The tool was communicated too broadly and lacked specificity of who would be responsible for reviewing green patients and advocating discharge. (8) The CDS was implemented insufficiently relying on regular checking which caused delays in awareness of newly green patients.

Modeling Medical Stability for Discharge Readiness

For a myriad of reasons, (correctly) estimating a patient's low risk may not decrease gLOS/LOS as there are other rate-limiting factors involved in discharge planning that were not addressed by this intervention. Being low-risk for COVID-19 adverse events is of drastically less importance in evaluating discharge readiness when a patient was primarily admitted for an unrelated reason. The system estimates risk for any hospitalized patient who tests positive for COVID-19—including asymptomatic patients. Among enrolled patients, only half ($n = 525$) had a primary diagnosis code consistent with COVID-19 (►**Supplementary Appendix Table C5**, available in the online version) which has a dilutive effect on statistical power where a trend of reduced gLOS is not detectable (►**Supplementary Appendix Table C6**, available in the online version).

The model was derived from data collected during a COVID-19 surge when the pressure to treat and discharge patients quickly was high. While the model may identify patients very likely to not deteriorate within 4 days, they may remain febrile and oxygen dependent. With surge pressure, these medically stable patients could be discharged earlier but as the surge waned and resources became more readily available, providers would prefer to keep these patients for observation, negating the effect of this intervention.

Medical stability is only one aspect of readiness for discharge alongside social readiness which was especially complex during COVID-19 surges when placement and home services were strained. With increasing cases and reduced bed capacity, hospitals would need to discharge otherwise stable patients on oxygen (as long as it did not exceed what could be administered by a home health provider). As the surge diminished, the clinical practice transitioned to discharge patients only once they could be weaned off oxygen. Similarly, some skilled nursing homes required patients to test negative before transfer causing patients to be held for days longer than medically required. These policies delay discharge and remove any potential for our intervention to have an effect.

Clinician Perceptions and Learning

Clinicians who received communication about the score may have chosen to not use it due to distrust or being unconvinced of any benefit over their clinical experience. Others may have been skeptical but tried it out only to have a poor initial experience, noting inaccuracies (perceived or accurate) as a reason to stop using it. Humans tend to be less forgiving of mistakes by algorithms than by other humans—referred to as algorithm aversion.[21] This skepticism may not be evident in the EHR metrics as users do not bother to remove the score from their patient lists and continue to use the summary report for other information.

Randomizing by patient exposes clinicians to both treatment and control arms which could also explain our results. As providers interact with intervention scores they will form a mental model of patients who have better predicted outcomes (lower scores). As a provider applies their mental model, it could spill over into the control arm, potentially diluting the intervention effect.[22] Contamination can be mitigated with double blinding—impractical with a CDS intervention—or alternative randomization strategies. Although randomizing by provider or clustered by department would mitigate potential for contamination, it is logistically complex in the inpatient setting. The pandemic environment introduced further challenges: for example, the group of providers assigned to each department or likely to treat COVID patients was dynamic (adjusting to the case burden) and unknown to prespecify a cluster or stepped-wedge randomization. Targeting a broad set of user roles in the inpatient setting also complicates randomization as some members of a patient's treatment team may be allocated treatment and others control. As the outcome measures were patient-based, the cleanest randomization strategy was patient-based which remains common in pragmatic trials.[23]

Outreach and CDS Integration

The new system was broadly communicated to departments across each hospital focusing on what the new risk score was, where to find it, and how to use it. However, these outreach efforts were limited to the very early period of the study when there was a barrage of COVID-19-specific communications that may have impaired user adoption given the amount of outreach typically employed for new EHR functionalities. The pandemic also complicated outreach channels as many newly recruited clinicians or those who did not specialize in hospital medicine were rapidly transitioned to COVID-19 departments. Use of the two CDS channels appears consistent over the study period (►**Fig. 2**) but preliminary evidence (four summary report displays per patient per week and one quarter of General Medicine hospitalists adding the patient list column) suggests these outreach efforts were insufficient to convince all individuals of the intervention's potential benefit. Repeated outreach may have created a larger effect.

The broad approach to outreach lacked specificity to empower a specific role of users to regularly check patient's scores and relay those to the larger care team. For example, care coordinators could review a large number of patients multiple times a day, identify green transitions, and prompt discharge. Early identification of medical or social barriers to

discharge could expedite resolution and discharge. By casting a wide net during outreach, the effect of the intervention may have been diluted.

The communicated use of the score as a tool to check periodically may be insufficiently timed to expedite discharge and reduce LOS. As the CDS is passive, displays of the score are inefficiently used on the majority who have not changed colors instead of highlighting patients who have newly transitioned to green. If an attending reviews their patient's scores each morning, they could miss up to 23.5 hours of green time—much larger than the target of 0.5 day reduction. The summary report metrics (►Fig. 2B) average four displays per patient per week suggesting the mean time between displays exceeds 24 hours. Greater adoption, more frequent use, or more targeted CDS could improve user awareness of recently green patients.

Lessons Learned for Future Work

We will build on this study with our future work in a variety of ways. We plan to appropriately randomize all future AI systems to ensure safety and assess efficacy in a robust manner and, if required, initiate a sundown protocol to remove unsafe or ineffective systems (as will be the case for this model). While we incorporated aspects of user-centered design with representation from attending physicians and clinical informaticians, we may have omitted vital perspectives from other groups, e.g., nursing, social work, and care management, and plan to engage a wider audience in scoping and design phases. We will be working on a set of protocols with our IT colleagues for how to design and distribute training materials for future systems. Finally, we plan to supplement RCT results with qualitative results to assess user engagement and perceptions after experience with the tool being live. Implementing and robustly evaluating AI systems is resource intensive; we plan to be stricter with future projects to require a clear audit trail of user engagement and action. If linkable process measures do not exist, we should build them to help promote adoption and evaluate intervention effects.

Limitations

Inpatient LOS is associated with quality but can be a challenging metric to evaluate due to the influence of non-clinical factors.[24,25] COVID-19 also has a highly variable disease course,[26,27] where many comorbidities are associated with worse outcomes.[28,29] This heterogeneity is evident by the wide range and skew of observed gLOS and LOS (►Table 2 and ►Supplementary Appendix Fig. C1 [available in the online version]). Power calculations aimed for a 0.5 day improvement in gLOS and estimated 500 patients per arm but as the surge pressure waned and care improved over time, median gLOS decreased (►Supplementary Appendix Table C1, available in the online version). The study was underpowered for smaller effect sizes expected with shorter median gLOS. A clinically meaningful effect may have been easier to detect in alternative patient outcomes.

Disentangling whether intervention adoption impacted the outcome is challenging due to limitations in measuring usage of the intervention. To limit clinician burden during this period, the CDS was designed to be passive and not require any documentation or acknowledgement. Introduced variability in how many users observed an individual participant's score could not be controlled for as the EHR metrics were anonymous (of both patient and provider). It was impossible to separate individuals, user types, or departments. Some high-utilizing users who frequently refresh small lists, for example, could have created positive trends among their patients but with no data source, we were unable to detect this. These metrics provide an unscaled measure of displays over time but otherwise fail as a process measure. We urge EHR vendors to develop more detailed, individualized measures of CDS use that can capture specific interactions such as clicks, scrolls, and hovers. For this work, usage data per-user, per-patient would have greatly helped to assess adoption, target outreach, and connect utilizers with their patients and outcomes. These metrics should not require additional clicks by users to log their review of a patient's score.

Conclusion

An AI-based CDS intervention displaying color-coded risk scores of COVID-19 related deterioration was safe, but did not reduce LOS compared to patients with their scores withheld. While AI predictions themselves are accurate, interventions must be tested rigorously in real-world clinical scenarios to ensure they improve outcomes. However, pragmatic challenges of adoption, poor process measures of agreement/acknowledgement, and provider learning complicate evaluations. Such complications are magnified in the setting of a health care crisis where clinical care processes are constantly evolving and clinicians are inundated with communication related to these changes. Better processes to enable studies of health technology and AI interventions are required to nurture and evaluate benefits.

Clinical Relevance Statement

The rapid emergence of a new disease, COVID-19, motivated the use of machine learning to help augment clinical care in a data-driven way when empirical evidence was lacking. Passive display of a patient's risk score was found to be safe but had no effect in reducing LOS. How providers use the score in their clinical practice remains unclear from this work as does how best to promote and measure use for future studies.

Multiple Choice Questions

1. By displaying color-coded risk scores to clinicians, which patient outcome was expected to improve?
   a. Oxygen requirements.
   b. Length of stay.
   c. Mortality.
   d. Time to intubation.

Correct Answer: The correct answer is option b. The hypothesis behind this study was that green-scored patients at low-risk of adverse events related to COVID-19 could be discharged sooner, reducing length of stay. Provider awareness of color-coded risk score would not affect the patient's physiology and oxygen requirements. Some providers may be more attentive to the patient's condition as a result of watching their scores which could feasibly reduce time to intubation or mortality for higher risk patients (colored orange or red) but this was not the expected use of the system.

2. Where were the risk scores displayed to clinicians?
   a. Chart review and nursing notes.
   b. Patient list and summary report.
   c. Progress notes and flowsheets.
   d. EHR banner and infection status.

Correct Answer: The correct answer is option b. In this study, the risk scores were displayed as a patient list column and via a COVID-19 summary report. The same information is presented in both channels as they draw from the standard epic functionality. Additional custom build would be required for display or documentation of scores during chart review and nursing notes, progress notes and flowsheets, or EHR banner and infection status.

Protection of Human and Animal Subjects

The study was performed in compliance with the World Medical Association Declaration of Helsinki on Ethical Principles for Medical Research Involving Human Subjects, and was reviewed by the Institutional Review Board of NYU Grossman School of Medicine.

Funding

Yindalon Aphinyanaphongs was partially supported by NIH 3UL1TR001445-05 and National Science Foundation award #1928614.

Conflict of Interest

None declared.

Acknowledgements

The authors thank Hardev (Dave) Randhawa for his help integrating the model into the EHR and George Redgrave for suggesting the EHR metrics.

References

1 Liu J, Li S, Liu J, et al. Longitudinal characteristics of lymphocyte responses and cytokine profiles in the peripheral blood of SARS-CoV-2 infected patients. EBioMedicine 2020;55:102763
2 Guan WJ, Ni ZY, Hu Y, et al; China Medical Treatment Expert Group for Covid-19. Clinical characteristics of coronavirus disease 2019 in China. N Engl J Med 2020;382(18):1708–1720
3 Chen R, Liang W, Jiang M, et al; Medical Treatment Expert Group for COVID-19. Risk factors of fatal outcome in hospitalized subjects with coronavirus disease 2019 from a nationwide analysis in China. Chest 2020;158(01):97–105
4 Huang C, Wang Y, Li X, et al. Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. Lancet 2020;395 (10223):497–506
5 Galloway JB, Norton S, Barker RD, et al. A clinical risk score to identify patients with COVID-19 at high risk of critical care admission or death: an observational cohort study. J Infect 2020;81(02):282–288
6 Ji M, Yuan L, Shen W, et al. A predictive model for disease progression in non-severely ill patients with coronavirus disease 2019. Eur Respir J 2020;56(01):2001234
7 Li Q, Zhang J, Ling Y, et al. A simple algorithm helps early identification of SARS-CoV-2 infection patients with severe progression tendency. Infection 2020;48(04): 577–584
8 Liang W, Liang H, Ou L, et al; China Medical Treatment Expert Group for COVID-19. Development and validation of a clinical risk score to predict the occurrence of critical illness in hospitalized patients with COVID-19. JAMA Intern Med 2020;180(08): 1081–1089
9 Yan L, Zhang HT, Goncalves J, et al. An interpretable mortality prediction model for COVID-19 patients. Nat Mach Intell 2020;2 (05):283–288
10 Singh K, Valley TS, Tang S, et al. Evaluating a widely implemented proprietary deterioration index model among hospitalized patients with COVID-19. Ann Am Thorac Soc 2021;18(07): 1129–1137
11 Wynants L, Van Calster B, Collins GS, et al. Prediction models for diagnosis and prognosis of covid-19: systematic review and critical appraisal. BMJ 2020;369:m1328
12 Harish K, Zhang B, Stella P, et al. Validation of parsimonious prognostic models for patients infected with COVID-19. BMJ Health Care Inform 2021;28(01):e100267
13 Razavian N, Major VJ, Sudarshan M, et al. A validated, real-time prediction model for favorable outcomes in hospitalized COVID-19 patients. NPJ Digit Med 2020;3:130
14 Wiens J, Saria S, Sendak M, et al. Do no harm: a roadmap for responsible machine learning for health care. Nat Med 2019;25 (09):1337–1340
15 Drysdale E, Dolatabadi E, Chivers C, et al. Implementing AI in Healthcare; 2019. Doi: 10.13140/RG.2.2.30793.70241
16 Begg C, Cho M, Eastwood S, et al. Improving the quality of reporting of randomized controlled trials. The CONSORT statement. JAMA 1996;276(08):637–639
17 Ogrinc G, Davies L, Goodman D, Batalden P, Davidoff F, Stevens D. Squire 2.0 (Standards for Quality Improvement Reporting Excellence): revised publication guidelines from a detailed consensus process. Am J Crit Care 2015;24(06): 466–473
18 Pinnock H, Barwick M, Carpenter CR, et al; StaRI Group. Standards for Reporting Implementation Studies (StaRI) Statement. BMJ 2017;356:i6795
19 Haynes RB, Del Fiol G, Michelson M, Iorio A. Context and approach in reporting evaluations of electronic health record-based implementation projects. Ann Intern Med 2020;172(Suppl 11): S73–S78
20 R Core Team. R: A Language and Environment for Statistical Computing. Published online 2018. Accessed August 31, 2021 at: https://www.R-project.org/
21 Dietvorst BJ, Simmons JP, Massey C. Algorithm aversion: people erroneously avoid algorithms after seeing them err. J Exp Psychol Gen 2015;144(01):114–126
22 Shcherbatykh I, Holbrook A, Thabane L, Dolovich LCOMPETE III investigators. Methodologic issues in health informatics trials: the complexities of complex interventions. J Am Med Inform Assoc 2008;15(05):575–580
23 Horwitz LI, Kuznetsova M, Jones SA. Creating a learning health system through rapid-cycle, randomized testing. N Engl J Med 2019;381(12):1175–1179
24 Rosenfeld LS, Goldmann F, Kaprio LA. Reasons for prolonged hospital stay; a study of need for hospital care. J Chronic Dis 1957;6(02):141–152

25 Brasel KJ, Lim HJ, Nirula R, Weigelt JA. Length of stay: an appropriate quality measure? Arch Surg 2007;142(05):461–465
26 Garibaldi BT, Fiksel J, Muschelli J, et al. Patient trajectories among persons hospitalized for COVID-19: a cohort study. Ann Intern Med 2021;174(01):33–41
27 Pereira NL, Ahmad F, Byku M, et al. COVID-19: understanding inter-individual variability and implications for precision medicine. Mayo Clin Proc 2021;96(02):446–463
28 Guan WJ, Liang WH, Zhao Y, et al; China Medical Treatment Expert Group for COVID-19. Comorbidity and its impact on 1590 patients with COVID-19 in China: a nationwide analysis. Eur Respir J 2020; 55(05):2000547
29 Sanyaolu A, Okorie C, Marinkovic A, et al. Comorbidity and its impact on patients with COVID-19. SN Compr Clin Med 2020;2 (08):1069–1076

