Early Identification of Hospitalized Patients with COVID-19 at Risk of Clinical Deterioration: A Multi-Site Study

Journal:	*BMJ*
Manuscript ID	BMJ-2021-068576
Article Type:	Original research
Date Submitted by the Author:	23-Sep-2021
Complete List of Authors:	Kamran, Fahad; University of Michigan College of Engineering Tang, Shengpu; University of Michigan College of Engineering, Division of Computer Science & Engineering Otles, Erkin; University of Michigan College of Engineering, Department of Industrial and Operations Engineering; University of Michigan Medical School McEvoy, Dustin; Mass General Brigham Inc Saleh, Sameh; The University of Texas Southwestern Medical Center Gong, Jen; University of California San Francisco Medical Center Li, Benjamin; University of Michigan Medical School Dutta, Sayon; Mass General Brigham Inc Liu, Xinran; University of California San Francisco Medical Center Medford, Richard; University of Texas Southwestern Medical Center at Dallas, Infectious Diseases; University of Texas Southwestern Medical Center at Dallas, Clinical Informatics Valley, Thomas; University of Michigan, Internal Medicine West, Lauren; Massachusetts General Hospital Singh, Karandeep; University of Michigan Medical School Blumberg, Seth; University of California San Francisco Medical Center Donnelly, John; University of Michigan Medical School Shenoy, Erica; Massachusetts General Hospital Ayanian, John; University of Michigan School of Medicine, Internal Medicine; University of Michigan Institute of Health Care Policy and Innovation Nallamothu, Brahmajee; Ann Arbor Health Services Research and Development Center of Excellence, Division of Cardiovascular Medicine; University of Michigan Medical School, Center for Health Outcomes and Policy Sjoding, Michael; University of Michigan, Pulmonary-Critical Care, Internal Medicine Wiens, Jenna; University of Michigan College of Engineering, Division of Computer Science & Engineering
Keywords:	Machine Learning, External Validation, Deterioration Index, COVID-19

Title: Early Identification of Hospitalized Patients with COVID-19 at Risk of Clinical Deterioration: A Multi-Site Study

Authors: Fahad Kamran, MS[1,*]; Shengpu Tang, MSE[1,*]; Erkin Otles, MEng[2,3]; Dustin S. McEvoy, BS[4]; Sameh N. Saleh, MD[5,6]; Jen Gong, PhD[7]; Benjamin Y. Li, BS[1,3]; Sayon Dutta, MD, MPH[4,8]; Xinran Liu, MD, MS, FAMIA[9]; Richard J. Medford, MD[5,6]; Thomas S. Valley, MD, MSc[10,11]; Lauren R. West, MPH[12]; Karandeep Singh, MD, MMSc[10,13]; Seth Blumberg, MD, PhD[9,14]; John P. Donnelly, PhD[10,13]; Erica Shenoy, MD, PhD[12,15,16]; John Z. Ayanian, MD, MPP[10,11]; Brahmajee K. Nallamothu, MD, MPH[10,11]; Michael W. Sjoding, MD, MSc[10,11,†]; Jenna Wiens, PhD[1,10,†]

*Co-first authors; †Co-senior authors

Affiliations:

1. Division of Computer Science and Engineering, University of Michigan College of Engineering, Ann Arbor, MI, USA.
2. Department of Industrial and Operations Engineering, University of Michigan College of Engineering, Ann Arbor, MI, USA.
3. Medical Scientist Training Program, University of Michigan Medical School, Ann Arbor, MI, USA.
4. Mass General Brigham Digital Health eCare, Somerville, MA, USA.
5. Department of Internal Medicine, University of Texas Southwestern Medical Center, Dallas, TX, USA.
6. Clinical Informatics Center, University of Texas Southwestern Medical Center, Dallas, TX, USA.
7. Center for Clinical Informatics and Improvement Research, University of California, San Francisco, CA, USA.
8. Department of Emergency Medicine, Massachusetts General Hospital and Harvard Medical School, Boston, MA, USA.
9. Division of Hospital Medicine, University of California, San Francisco, San Francisco, CA, USA.
10. Institute for Healthcare Policy and Innovation, University of Michigan, Ann Arbor, MI, USA.
11. Department of Internal Medicine, University of Michigan Medical School, Ann Arbor, MI, USA.
12. Infection Control Unit, Massachusetts General Hospital, Boston, MA, USA.
13. Department of Learning Health Sciences, University of Michigan Medical School, Ann Arbor, MI, USA.
14. Francis I. Proctor Foundation, University of California, San Francisco, San Francisco, CA, USA.
15. Department of Medicine, Harvard Medical School, Boston, MA, USA.
16. Division of Infectious Diseases, Massachusetts General Hospital, Boston, MA, USA.

Correspondence to: Jenna Wiens, PhD, Division of Computer Science and Engineering, Department of Electrical Engineering and Computer Science, University of Michigan, 2260 Hayward Street, Ann Arbor, MI 48109, USA (wiensj@umich.edu).

Article Type: BMJ - Research
Word count: 3262
Figures and Tables: 1 Table + 4 Figures

Keywords: Machine Learning, External Validation, Deterioration Index, COVID-19

ABSTRACT (361 words)

Objectives: Simple, transferable and accurate methods for patient risk stratification are needed to better plan and allocate resources, as highlighted by the strain on hospitals created by the COVID-19 pandemic. Using a novel paradigm of model development and code sharing, we sought to create a machine learning model from electronic health record (EHR) data that can accurately predict patient deterioration across institutions.

Design, Setting, Participants: In a retrospective cohort study, hospitalized adults with respiratory distress at one institution from 2015-2021 were used for model training and internal validation. External validation was conducted on patients hospitalized with COVID-19 during 2020-2021 at 12 additional US medical centers.

Main Outcomes Measure: On the internal development cohort, an ensemble of linear models was trained to predict a composite outcome of in-hospital mortality and three events indicating need for ICU-level therapies: 1) mechanical ventilation, 2) heated high-flow nasal cannula and 3) intravenous vasopressors, based on 9 clinical and demographic variables selected from 2,686 variables available in the EHR. Internal and external validation performance was measured using the area under the receiver operating characteristic curve (AUROC) and the expected calibration error (ECE), i.e., the difference between predicted risk and actual risk. Potential bed-day savings were estimated by calculating how many days per patient the hospitals could save if low-risk patients identified by the model were discharged early.

Results: A total of 9,291 COVID-19 hospitalizations at 13 medical centers were used for model validation, of which 1,510 (16.3%) experienced the primary outcome. On the internal validation cohort, the model achieved an AUROC of 0.80 (95% CI: 0.77, 0.84) and an ECE of 0.01 (95%

CI: 0.00, 0.02). Performance was consistent in the 12 external medical centers (AUROC range: 0.77-0.84), across demographic subgroups of sex, age, race, and ethnicity (AUROC range: 0.78-0.84), and across quarters (AUROC range: 0.73-0.83). Using the model to triage low-risk patients could potentially save up to 7.8 bed-days per early discharge.

Conclusion: A deterioration model developed rapidly in response to the pandemic at a single hospital was applied externally without sharing data and generalized across multiple medical centers, demographic subgroups and time periods, demonstrating its potential as a tool for use in optimizing healthcare resources.

INTRODUCTION

Risk stratification models that provide advance warning of patients at high-risk of deterioration during hospitalization could help clinical care teams manage resources, including interventions, hospital beds and staffing.[1,2] For example, knowing how many and which patients will require ventilators could prompt hospitals to increase ventilator supply while care teams start to allocate ventilators to patients most in need.[3] Beyond identifying high-risk patients, such models could also help identify low-risk patients as candidates for early discharge, potentially freeing up hospital resources.[4–7]

Despite the potential use of risk stratification models in resource allocation, few successful examples exist. Most notably, strong generalization performance, i.e., how well a model will perform across different patient populations, is fundamental to realizing the potential benefits of risk models in clinical care. Yet generalization performance is often entirely overlooked when developing and validating predictive models in healthcare.[8–14] For example, recent work found that only 5% of articles on predictive modeling in PubMed mention external validation in either the title or the abstract.[9] This is due, in part, to the fact that most approaches to external validation require data-sharing agreements.[15–18] In the small fraction of cases in which data sharing agreements have been successfully established, validation was either limited in scope[19–21] (e.g., focused on a single geographical region) or the model performed poorly once applied to a population that differed from the development cohort.[22,23] Thus, there is a critical need for an accurate, simple and open-source method for patient risk stratification that generalizes across hospitals and patient populations.

In this study, we develop and validate an open-source patient deterioration model, Michigan Critical Care Utilization and Risk Evaluation System (M-CURES), using routinely available data extracted from electronic health records (EHR). We externally validate this risk model across

multiple dimensions, while preserving data privacy and forgoing the need for data sharing across healthcare institutions. To evaluate the effectiveness of the model in settings where risk stratification could be highly beneficial, we focus on patients hospitalized with COVID-19 from 13 US medical centers. COVID-19 represents an important case study given increases in hospitalizations during the COVID-19 pandemic have strained hospital resources on a global scale;[24–26] some hospitals have been forced to cancel up to 85% of elective surgical procedures to free up resources.[27,28] We hypothesized that a simple model based on a handful of variables would generalize across diverse patient cohorts.

METHODS

This study was approved by the institutional review boards of all participating sites, with a waiver of informed consent. Additional methodological details can be found in the **Supplement**.

Study Cohorts

Development Cohort. The model was trained on adult (18 years and older) patient hospitalizations at Michigan Medicine, the academic medical center of the University of Michigan, during the 5-year period from January 1, 2015 to December 31, 2019. All hospitalizations with respiratory distress, i.e., those admitted through the emergency department who received supplemental oxygen support, were included. Hospitalizations that met the outcome (described below) prior to or at the time of receiving supplemental oxygen were excluded.

Internal Validation Cohort. The model was internally validated on adult patient hospitalizations at Michigan Medicine from March 1, 2020 to February 28, 2021 who required supplemental oxygen and were diagnosed with COVID-19. To identify COVID-19 hospitalizations from retrospective data, we included hospitalizations with either 1) a positive laboratory test or 2) a

recorded ICD-10 code for COVID-19 without a negative laboratory test. A randomly selected subset of 100 hospitalizations were used for variable selection and excluded from evaluation.

External Validation Cohorts. The external validation cohorts included adult patient hospitalizations at 12 external medical centers from March 1, 2020 to February 28, 2021 who required supplemental oxygen and were diagnosed with COVID-19. Inclusion criteria were similar to those used for the internal validation cohort (**eMethods 1** in **Supplement**).

In alphabetical order, the external healthcare systems were Mass General Brigham (MGB), the University of California San Francisco Medical Center, and University of Texas Southwestern Medical Center. MGB included 10 hospitals: Brigham and Women's Faulkner Hospital, Brigham and Women's Hospital, Cooley Dickinson Hospital, Martha's Vineyard Hospital, Massachusetts General Hospital, McLean Hospital, Nantucket Cottage Hospital, Newton-Wellesley Hospital, North Shore Medical Center, and Wentworth-Douglass Hospital. Six sites with fewer than 100 cases that met the primary outcome were combined into a single cohort when performing evaluation, resulting in a total of 7 external validation cohorts. These medical centers represent both large academic medical centers and small to mid-size community hospitals in regions geographically distinct from the development institution (Midwest), including the Northeast, West, and South regions of the US. Institution-specific results were anonymized.

Cohort Comparison. We compared the internal validation cohort to the development cohort and to each of the external validation cohorts across demographic characteristics and outcomes, using chi-squared tests for homogeneity with a Bonferroni correction for multiple comparisons, at a significance level of $\alpha=0.001$.

Outcome

The model was trained to predict a composite outcome of clinical deterioration, defined as either in-hospital mortality or the need for intensive care unit (ICU)-level therapies, including the receipt of invasive mechanical ventilation, heated high-flow nasal cannula, or intravenous vasopressors. The outcome time was defined as the earliest (if any) of these events within the first five days of hospitalization. Additional implementation details are described in **eMethods 1** in the **Supplement**.

Model Development & Evaluation

Variable Selection and Feature Engineering. A model based on data extracted from the EHR was developed to predict the primary outcome every 4-hours (at set time points; see **eFigure 1** in **Supplement**). Clinical knowledge and data-driven feature selection was used to reduce the input space from 2,686 EHR variables (including demographics, laboratory results, and data recorded in nursing flowsheets) to 9 variables. First, variables with the potential to be spuriously correlated with the outcome were removed based on clinical expertise.[29] In addition, variables that relied on existing deterioration indices or composite scores (e.g., the SOFA score[30]) were removed, due to the potential for inconsistencies or lack of availability across healthcare systems. Then, using 100 randomly selected hospitalizations from the internal validation cohort, permutation importance[31,32] and forward selection[33] were used to further reduce the variable set (**eMethods 2** in **Supplement**). The final 9 variables included: age, respiratory rate, oxygen saturation, oxygen flow rate, pulse oximetry type (e.g., continuous, intermittent), head of bed position (e.g., at 30 degrees), patient position when blood pressure was measured (e.g., standing, sitting, lying), venous blood gas pH, and arterial blood gas pCO2. FIDDLE,[34] an open-source preprocessing pipeline for structured EHR data, was used to map the 9 data elements to 88 binary features (each with a value in {0,1}) describing every 4-hour window. The features were used as input to the machine learning model, and included summary information

about each variable (e.g., the minimum/maximum/mean respiratory rate within a window) and indicators for missingness (e.g., whether respiratory rate was measured within a window).

Model training. An ensemble of regularized logistic regression models was trained to map patient features from each 4-hour window to an estimate of clinical deterioration risk. From the development cohort, a single 4-hour window was randomly sampled for each hospitalization to train a logistic regression model. For hospitalizations in which the outcome occurred, only windows prior to the one before the outcome were used. The process was repeated 500 times, leading to an ensemble of 500 models, whose outputs were averaged to create a final prediction. Models were trained to predict whether a hospitalization would experience the primary outcome within five days of hospitalization. Additional details are described in **eMethods 2** in **Supplement**.

Internal Validation. Model discriminative performance was measured using the area under the receiver operating characteristics curve (AUROC) and the area under the precision-recall curve (AUPR). Models were evaluated from the first full window of data, with model predictions beginning in the window with a hospitalization's first vital signs being recorded. The model aims to support clinical decision making prospectively during which a risk score is recomputed every 4 hours and the care team decides whether or not to intervene once the hospitalization reaches a certain score. For this reason, all evaluations were performed at the hospitalization-level, rather than the 4-hour window-level (**eMethods 2** in **Supplement**). Model calibration was assessed using reliability curves and expected calibration error (ECE) based on quintiles of predicted risk, i.e., the average absolute difference between predicted risk and observed risk.[35,36] Calibration was evaluated at the window-level to measure how well each prediction aligns with absolute risk. As a baseline, the model was compared to a common proprietary model, the Epic Deterioration Index,[37] in the internal validation cohort.

External Validation. Research teams at each collaborating institution applied the inclusion/exclusion criteria locally to identify an external validation cohort at their institution. Once cohorts were identified, local teams extracted the 9 clinical and demographic variables described above and saved their data to match a requested format that would allow application of identical preprocessing. In addition, teams applied the outcome definition to determine which hospitalizations experienced clinical deterioration. After preprocessing, each team independently applied the same model and evaluation code (**eMethods 2** in **Supplement**) and reported results as summary statistics. As in the internal validation, the model was evaluated in terms of both discriminative and calibration performance in each external cohort. Internal performance and external performance were compared using a bootstrap resampling test by computing 95% confidence intervals (CIs) of the difference in performance, adjusted by Bonferroni correction.

Assessing Model Generalizability Across Time and Demographic Subgroups. To further evaluate model performance across time, the AUROC and AUPR scores were measured for every quarter (three-month periods) between March 2020 and February 2021 within each validation cohort. Performance was also evaluated across different demographic subgroups as the mean (and standard deviation) of AUROC scores across cohorts for different subgroups of sex, age, race, and ethnicity (categorizations in **eMethods 2** in **Supplement**). Within each cohort, subgroup performance was compared to overall performance using the same bootstrap resampling approach described above.

Identifying Low-risk Patients. To further examine how the model might be applied in hospitals for resource allocation, the model was evaluated for its ability to identify hospitalizations in which the patient who did not develop the outcome after 48 hours of observation. For each

validation cohort, the percentage of hospitalizations correctly flagged as low risk was calculated for a negative predictive value (NPV) greater than or equal to 95% (i.e., of the hospitalizations flagged as low risk, 5% or fewer met the outcome). From this estimate, the number of bed days that could potentially be saved if these patients had been discharged at 48 hours was reported (**eMethods 2** in **Supplement**).

Implementation Details & Code Sharing Statement

All analyses were performed in Python 3.5.2[38] using the numpy,[39] pandas,[40,41] sklearn[42] packages. Code for data preprocessing and model evaluation were packaged, and each institution ran the same pipeline locally and independently. All code and documentation are available online at https://github.com/MLD3/M-CURES, (in a repository that will be made public upon publication) so that other institutions can validate and use the model.

RESULTS

The development cohort included 35,040 hospitalizations between 2015 and 2019 at a single institution, of which 3,757 (10.7%) experienced the primary outcome (**eTable 1** in **Supplement**). The internal validation cohort included 956 hospitalizations in which the patient had COVID-19, of which 206 (21.6%) experienced the primary outcome (**Table 1**). Compared to the development cohort, hospitalizations in the internal validation cohort were similar in age and sex but were more likely to be Black (19.6% vs. 11.3%) (**eTable 1** in **Supplement**). Combined, the external validation cohorts consisted of 8,335 hospitalizations, of which 1,304 (15.6%) experienced the primary outcome. All external validation cohorts differed from the internal validation cohort in at least one demographic dimension (sex, age, race, and ethnicity) (**Table 1**; **eTable 2** in **Supplement**). For example, the proportions of Hispanic or Latino patients were significantly higher, ranging 13.5%-29.0% vs. 3.6%; in four external cohorts there was a significantly larger proportion of very elderly patients (>85 yrs), with one institution skewing

Table 1. Characteristics of internal and external validation cohorts. We included all adult hospitalizations with a COVID-19 diagnosis between March 1, 2020 and February 28, 2021, from an internal validation cohort (MM) and 7 external validation cohorts (A-G) pertaining to 12 medical centers. Characteristics of the development cohort can be found in **eTable 1** in **Supplement**.

Institution	MM	A	B	C	D	E	F	G
Number of patients	887	2161	1252	1180	1009	909	747	555
Number of hospitalizations	956	2320	1320	1256	1073	965	794	607
Median age in years [IQR]	64 [52–75]	63 [50-76]	62 [50–73]	68 [56-79]	65 [53-76]	69 [58-80]	73 [59-84]	62 [48–75]
Age Group (%)								
[18, 25]	<25	52 (2.2)	<25	<25	<25	<25	<25	<25
(25, 45]	129 (13.5)	398 (17.2)	225 (17.1)	159 (12.7)	159 (14.8)	77 (8.0)	74 (9.3)	114 (18.8)
(45, 65]	374 (39.1)	800 (34.5)	518 (39.2)	380 (30.3)	358 (33.4)	327 (33.9)	204 (25.7)	215 (35.4)
(65, 85]	365 (38.2)	873 (37.6)	497 (37.7)	539 (42.9)	435 (40.5)	412 (42.7)	331 (41.7)	184 (30.3)
>85	70 (7.3)	197 (8.5)	57 (4.3)	159 (12.7)	97 (9.0)	145 (15.0)	177 (22.3)	74 (12.2)
Sex (%)								
Female	420 (43.9)	993 (42.8)	612 (46.3)	564 (44.9)	533 (49.7)	445 (46.1)	363 (45.7)	313 (51.6)
Male	536 (56.1)	1327 (57.2)	709 (53.7)	692 (55.1)	540 (50.3)	520 (53.9)	431 (54.3)	294 (48.4)
Race (%)								
White	649 (67.9)	1364 (58.8)	733 (55.6)	935 (74.4)	589 (54.9)	636 (65.9)	584 (73.6)	214 (35.3)
Black	187 (19.6)	190 (8.2)	332 (25.2)	123 (9.8)	234 (21.8)	135 (14.0)	49 (6.2)	62 (10.2)
Asian	30 (3.1)	80 (3.4)	29 (2.2)	51 (4.1)	39 (3.6)	<25	39 (4.9)	135 (22.2)
Other/Unknown	90 (9.4)	686 (29.6)	226 (17.1)	147 (11.7)	211 (19.7)	168 (17.4)	122 (15.4)	196 (32.3)
Ethnicity (%)								
Hispanic or Latino	34 (3.6)	587 (25.3)	379 (28.7)	350 (27.9)	210 (19.6)	138 (14.3)	107 (13.5)	176 (29.0)
Not Hispanic or Latino	883 (92.4)	1569 (67.6)	915 (69.3)	875 (69.7)	841 (78.4)	783 (81.1)	637 (80.2)	414 (68.2)
Other/Unknown	39 (4.1)	164 (7.1)	26 (1.8)	31 (2.5)	<25	44 (4.6)	50 (6.3)	<25
Median LOS in hours [IQR]	138 [83-261]	160 [95-284]	141 [96-257]	136 [93-235]	167 [100-287]	143 [92-234]	154 [95-256]	183 [113-324]
Outcome ever (%)								
Death	60 (6.3)	197 (8.5)	108 (8.2)	125 (10.0)	96 (8.9)	93 (9.6)	123 (15.5)	42 (6.9)
MV	98 (10.3)	259 (11.2)	142 (10.7)	135 (10.7)	116 (10.8)	69 (7.2)	69 (8.7)	52 (8.6)
IV	87 (9.1)	299 (12.9)	152 (11.5)	139 (11.1)	125 (11.6)	65 (6.7)	74 (9.3)	70 (11.5)
HHFNC	218 (22.4)	132 (5.7)	263 (19.9)	121 (9.6)	95 (8.9)	99 (10.3)	106 (13.4)	101 (16.6)
Primary Outcome <= 5 days	206 (21.6)	311 (13.4)	249 (18.8)	206 (16.4)	155 (14.4)	136 (14.1)	155 (19.5)	92 (15.2)
Reason for primary outcome (% of outcomes)								
Death	5 (2.4)	34 (10.9)	4 (1.6)	21 (10.2)	16 (10.3)	25 (18.4)	37 (23.9)	2 (2.2)
MV	20 (9.7)	89 (28.6)	25 (10.0)	52 (25.2)	52 (33.5)	22 (16.2)	18 (11.6)	8 (8.7)
IV	9 (4.4)	95 (30.5)	18 (7.2)	33 (16.0)	26 (16.8)	10 (7.4)	21 (13.5)	16 (17.4)
HHFNC	172 (83.5)	93 (29.9)	202 (81.1)	100 (48.5)	61 (39.4)	79 (58.1)	79 (51.0)	66 (71.7)

Acronyms: IQR, interquartile range; LOS, Length-of-Stay; MV, Mechanical Ventilation; IV, Intravenous Vasopressors, HHFNC, Heated High-Flow Nasal Cannula.

much older (22.3% vs. 7.3%). Externally, primary outcome rates varied from 13.4% to 19.5%. In addition, the reason for meeting the primary outcome varied significantly across hospitals (**eTable 3** in **Supplement**).

The parameters of the final learned model are visualized in **eFigure 2** in **Supplement**. The model demonstrated good overall performance in both internal and external validation. Applied to the internal validation cohort, it substantially outperformed the Epic Deterioration Index, achieving an AUROC of 0.80 (95% CI: 0.77, 0.84) vs. 0.66 (95% CI: 0.62, 0.70), AUPR of 0.55 (95% CI: 0.48, 0.63) vs. 0.31 (95% CI: 0.26, 0.36) and ECE of 0.01 (95% CI: 0.00, 0.02) vs. 0.31 (95% CI: 0.30, 0.32) (**eFigure 3** in **Supplement**). External validation resulted in similar performance, with AUROC ranging 0.77-0.84, AUPR ranging 0.34-0.57, and ECE ranging 0.02-0.04 (**Figure 1**). The AUROC across external institutions did not differ significantly from the internal validation AUROC (**eTable 4** in **Supplement**) and had an average of 0.81.

Across time (**Figure 2**), the model performed consistently in all validation cohorts throughout the 4 quarters, with AUROC > 0.7 and AUPR > 0.2 in most cases. The major exception was during Jun-Aug 2020, where compared to the overall performance of each cohort, two cohorts had a drop in AUROC (from 0.79 to 0.57 and from 0.77 to 0.58) and one cohort had a drop in AUPR (from 0.42 to 0.17), but the differences were not statistically significant (**eTable 5** in **Supplement**). Across demographic subgroups, the model displayed consistent discriminative performance in terms of AUROC (**Figure 3**); subgroup performance did not vary significantly from the overall performance when evaluated within specific sex, age, race, ethnicity subpopulation (**eTable 6** in **Supplement**). In one external cohort, the model performed significantly better on Asian patients compared to White patients (**eTable 7** in **Supplement**).

In terms of resource allocation and planning, the model was able to accurately identify low-risk patients after 48 hours of observation in both the internal and external cohorts. At best, the model could correctly triage up to 41.6% of low-risk COVID-19 hospitalizations to lower acuity care, potentially saving 5.2 bed days for each early discharge (**Figure 4**). The model achieved this performance level while maintaining a NPV of at least 95%, i.e., of the hospitalizations flagged as low risk, 5% or fewer met the outcome.

DISCUSSION

Accurate predictions of patient deterioration can assist clinicians in risk assessment over a patient's hospitalization by identifying who might be in need of ICU-level care in advance of deterioration.[43–45] In surge scenarios, hospitals might use predictions to manage limited resources (e.g., beds) by triaging low-risk individuals to lower-acuity care. To this end, we developed an open-source patient risk stratification model that uses 9 routinely collected demographic and clinical variables from the EHR for prediction of clinical deterioration. Compared to previous deterioration indices that have failed to generalize across multiple patient cohorts,[22,46] the model achieved good performance when externally validated in 12 different medical centers. External validation can highlight blind spots when the validation cohort differs substantially from the development cohort, including clinical conditions (e.g., COVID-19 is a new disease), demographics (e.g., race and ethnicity), clinical workflows, and hospital sizes. The model's strong generalizability may be attributed in part to a separate but related development cohort for training, the clinician-informed data-driven approach to feature selection and a rigorous approach to internal validation.

We also evaluated performance on specific demographic subgroups (based on age, sex, race, and ethnicity) and across time.[47,48] Ensuring consistent performance across demographic subgroups can help mitigate biases against certain vulnerable populations.[49–51] Despite an

underrepresentation of Hispanic/Latino patients in the development institution relative to the external cohorts, model performance in this subgroup was consistent with non-Hispanic/Latino performance. At several points throughout the pandemic, changes in the patient population presenting with severe disease and changes to clinical workflows, treatments, and outcomes could have a substantial impact on how risk models may perform.[52–57] These changes may have resulted in a modest model performance decline at two sites in the summer of 2020, as specified in the results. However, performance then stabilized in the fall and winter surges, which may indicate a convergence in treatment for COVID-19.

Unlike previous work on the external validation of patient risk stratification models,[21] our approach did not rely on sharing data across multiple sources. Instead, we developed the model using data from a single institution and then shared code with external institutions who then applied the model to their data using their own computing platforms. This approach has many benefits. Sharing and aggregating data containing protected health information (e.g., dates) from 12 healthcare systems into a single repository would have required extensive data use agreements and additional computational infrastructure and added substantial time delays to model evaluation. Maintaining patient data internally further mitigates the potential risk of data access breaches. In addition to distributing the workload and evaluation process, this approach introduced fewer errors because each team was most familiar with their own data and thus less likely to make incorrect assumptions when identifying the cohort, model variables, and outcomes.

The success of this paradigm relied on several design decisions early in the process as well as continued collaboration throughout. First, the number of variables used by the model was limited, ensuring that all variables could be reliably identified and validated at each institution. Beyond model inputs, it was equally crucial to validate inclusion/exclusion criteria and outcome

definitions. To this end, we worked closely with both clinicians and informaticists from each institution to establish accurate definitions. Finally, we developed a code workflow with common input/output formats and shared detailed documentation. This in turn allowed for quick iteration among institutions, facilitating debugging.

The current analysis should be interpreted in the context of its study design. Importantly, a single EHR vendor (Epic Systems) was used across all medical centers. This commonality between institutions facilitated model implementation. Despite a common EHR vendor, however, local implementation of each EHR system requires local institutional knowledge, which was a feature of our multi-site team approach. To further ensure the model can generalize to more institutions, researchers should focus on validating the model in healthcare systems utilizing different EHR systems. Moreover, the model was developed and validated on adults with respiratory distress and a diagnosis of COVID-19 in the US. The model may or may not apply to individuals with respiratory distress without a COVID-19 diagnosis, or in countries with fewer healthcare resources. Furthermore, when estimating 'potential bed days saved' resulting from triaging low-risk patients, we assumed that those patients could be safely discharged at 48 hours. However, there might be other reasons that a patient may need to remain in the hospital, preventing early discharge. Finally, the composite outcome we considered was developed early in the pandemic based on clinical workflows and treatments at the time. As treatments evolve, outcome definitions might change which could affect model performance. Without implementation into clinical practice, it is unknown whether the use of such a model has an impact on clinical or operational outcomes such as early discharge planning.

This study represents an important step toward building and externally validating models for identifying individuals at both high and low risk of deteriorating within their hospital stay. The model transferred across a variety of institutions, subgroups and time periods. Our method for

external validation alleviates potential concerns surrounding patient privacy by forgoing the need for data sharing, while still allowing for realistic and accurate evaluations of a model within different patient settings. Thus, the implications are two-fold; the work here can help develop models for patient deterioration within a single institution and can promote external validation and multi-center collaborations without the need for data sharing agreements.

Summary Box

What is already known on this topic?

- Risk stratification models can augment clinical care and help hospitals better plan and allocate resources in healthcare settings.
- A useful risk stratification model should generalize across different patient populations, though generalization is often overlooked when developing models due to the difficulty of sharing patient data for external validation.
- Models that have been externally validated have failed to generalize to populations that differed from the cohort on which the models were built.

What this study adds

- Our study presents a paradigm for model development and external validation without the need for data sharing, while still allowing for quick and thorough evaluations of a model within different patient populations.
- Our study suggests the use of data-driven feature selection combined with clinical judgement can help identify meaningful features that allow the model to generalize across a variety of patient settings.

ETHICS STATEMENTS

This study was approved by the institutional review boards of all participating sites (University of Michigan | Michigan Medicine HUM00179831, Mass General Brigham 2012P002359, University of Texas Southwestern Medical Center STU-2020-0922, University of California San Francisco 20-31825), with a waiver of informed consent.

DATA AVAILABILITY STATEMENT

To guarantee the confidentiality of personal and health information, only the authors have had access to the data during the study in accordance with the relevant license agreements. The full model (including model coefficients and supporting code) will be released online via a public repository.

ACKNOWLEDGEMENTS

The authors would like to thank the Data Office for Clinical & Translational Research at the University of Michigan for their help in data extraction and curation. In addition, the authors would like to thank Melissa Wei, Ian Fox, Jeeheh Oh, Harry Rubin-Falcone, Donna Tjandra, Sarah Jabbour, Jiaxuan Wang, and Meera Krishnamoorthy, for helpful discussions during early iterations of this work.

FOOTNOTES

Contributors: Ayanian, Nallamothu, Sjoding, Wiens conceptualized the study. Kamran, Tang, Otles, McEvoy, Saleh, Gong, Li, Dutta, Liu, Medford, Valley, West, Singh, Blumberg, Donnelly, Sjoding, Wiens contributed to acquisition, analysis, or interpretation of data. Kamran, Tang, McEvoy, Saleh, Gong, Sjoding, Wiens had access to study data pertaining to their respective institutions and take responsibility for the integrity of the data and the accuracy of the data analysis. Kamran, Tang, Otles drafted the manuscript. Kamran, Tang, Otles, McEvoy, Saleh,

Gong, Li, Dutta, Liu, Medford, Valley, West, Singh, Blumberg, Donnelly, Shenoy, Ayanian, Nallamothu, Sjoding, Wiens provided critical revision of the manuscript for important intellectual content. Nallamothu, Sjoding, Wiens supervised the conduct of this study. The corresponding author attests that all listed authors meet authorship criteria and that no others meeting the criteria have been omitted.

Funding: This work was supported by the National Science Foundation (award IIS-1553146 to Wiens), by the National Institutes of Health -National Library of Medicine (grant R01LM013325 to Wiens and Sjoding) -National Heart, Lung, & Blood Institute (grant K23HL140165 to Valley), by the Agency for Healthcare Research and Quality (grant R01HS028038 Valley), by the Centers for Disease Control and Prevention -National Center for Emerging and Zoonotic Infectious Diseases (grant U01CK000590 to Blumberg), and by Precision Health and the Institute for Healthcare Policy and Innovation at the University of Michigan. The funding sources had no role in the design and conduct of the study; collection, management, analysis, and interpretation of the data; preparation, review, or approval of the manuscript; and decision to submit the manuscript for publication. The views and conclusions in this document are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of the U.S. Government, the Department of Veterans Affairs, the National Science Foundation, the National Institutes of Health, the Agency for Healthcare Research and Quality, or the Centers for Disease Control and Prevention.

Competing Interests: Medford reported receiving funding from the Sergey Brin Family Foundation during the conduct of the study. Singh reported receiving grants from Blue Cross Blue Shield of Michigan and Teva Pharmaceuticals during the conduct of the study. Wiens reported receiving funding from Cisco Systems during the study. Dutta reported receiving funding from Agency for Healthcare Research and Quality as well as the CRICO/Risk

Management Fund during the conduct of the study. Singh reported receiving grant funding from Blue Cross Blue Shield of Michigan and Teva Pharmaceuticals during the conduct of the study. No other disclosures were reported.

Transparency Declaration: The lead author affirms that the manuscript is an honest, accurate, and transparent account of the study being reported; that no important aspects of the study have been omitted; and that any discrepancies from the study as originally planned (and, if relevant, registered) have been explained.

Patient and Public Involvement: Patients or the public were not involved in the design, or conduct, or reporting, or dissemination plans of our research.

FIGURE LEGENDS

Figure 1. Model performance across the internal and external validation cohorts. We measure discriminative performance in (A) ROC curves and (B) PR curves. Model calibration is shown in (C) Reliability plots based on quintiles of predicted scores. Results with 95% confidence intervals are summarized in (D). The internal validation cohort at Michigan Medicine (MM) is bolded, while the external validation cohorts A-G are shown in different colors. Overall, discriminative performance and calibration performance was good across institutions. The AUPR varied most in part due to variation in outcome rates.

Figure 2. Model discriminative performance (AUROC and AUPR scores) over the year broken down by quarter. The table denotes the legend and the number of hospitalizations included within each cohort in each quarter along with the percentage that met the outcome (in parentheses). The discriminative performance varied most in the second quarter during which there were the fewest number of patients who met the primary outcome. The AUROC across institutions varied little by the fourth quarter or third wave of the pandemic.

Figure 3. Model discriminative performance (AUROC scores) evaluated across demographic subgroups. Values are macro-average performance across institutions (error bars are ± one standard deviation). Across subgroups the AUROC did not vary significantly from the overall performance.

Figure 4. The model can be used to identify potential candidates for early discharge after 48 hours of observation. Using a decision threshold that achieves a negative predictive value of greater or equal to 95%, both the proportion of patients that could be discharged early (top) and the bedtime savings (in days), normalized by the number of correctly discharged

hospitalizations at each institution (bottom), are depicted. Results are computed over 1000 bootstrap replications.

REFERENCES

1 Shah NH, Milstein A, Bagley SC PhD. Making Machine Learning Models Clinically Useful. *JAMA* 2019;**322**:1351–2. doi:10.1001/jama.2019.10306

2 Peterson ED. Machine Learning, Predictive Analytics, and Clinical Practice: Can the Past Inform the Present? *JAMA* 2019;**322**:2283–4. doi:10.1001/jama.2019.17831

3 White DB, Lo B. A Framework for Rationing Ventilators and Critical Care Beds During the COVID-19 Pandemic. *JAMA* 2020;**323**:1773–4. doi:10.1001/jama.2020.5046

4 Coley CM, Li Y-H, Medsger AR, *et al.* Preferences for Home vs Hospital Care Among Low-Risk Patients With Community-Acquired Pneumonia. *Arch Intern Med* 1996;**156**:1565–71. doi:10.1001/archinte.1996.00440130115012

5 Page K, Barnett AG, Graves N. What is a hospital bed day worth? A contingent valuation study of hospital Chief Executive Officers. *BMC Health Serv Res* 2017;**17**:137. doi:10.1186/s12913-017-2079-5

6 Razavian N, Major VJ, Sudarshan M, *et al.* A validated, real-time prediction model for favorable outcomes in hospitalized COVID-19 patients. *npj Digital Medicine* 2020;**3**:130. doi:10.1038/s41746-020-00343-x

7 Pericàs JM, Cucchiari D, Torrallardona-Murphy O, *et al.* Hospital at home for the management of COVID-19: preliminary experience with 63 patients. *Infection* 2021;**49**:327–32. doi:10.1007/s15010-020-01527-z

8 Habib AR, Lin AL, Grant RW. The epic sepsis model falls short-the importance of external validation. *JAMA Intern Med* 2021;**181**:1040–1. doi:10.1001/jamainternmed.2021.3333

9 Ramspek CL, Jager KJ, Dekker FW, *et al.* External validation of prognostic models: what, why, how, when and where? *Clin Kidney J* 2021;**14**:49–58. doi:10.1093/ckj/sfaa188

10 Siontis GCM, Tzoulaki I, Castaldi PJ, *et al.* External validation of new risk prediction models is infrequent and reveals worse prognostic discrimination. *J Clin Epidemiol* 2014;**68**:25–34. doi:10.1016/j.jclinepi.2014.09.007

11 Steyerberg EW, Harrell FE Jr. Prediction models need appropriate internal, internal-external, and external validation. *J Clin Epidemiol* 2016;**69**:245–7. doi:10.1016/j.jclinepi.2015.04.005

12 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**. doi:10.1136/bmj.m1328

13 Cummings BC, Ansari S, Motyka JR, *et al.* Predicting Intensive Care Transfers and Other Unforeseen Events: Analytic Model Validation Study and Comparison to Existing Methods. *JMIR Med Inform* 2021;**9**:e25066. doi:10.2196/25066

14 Shamout FE, Shen Y, Wu N, *et al.* An artificial intelligence system for predicting the deterioration of COVID-19 patients in the emergency department. *npj Digital Medicine* 2021;**4**:80. doi:10.1038/s41746-021-00453-0

15 Price WN, Cohen IG. Privacy in the age of medical big data. *Nat Med* 2019;**25**:37–43. doi:10.1038/s41591-018-0272-7

16 Panch T, Mattie H, Celi LA. The “inconvenient truth” about AI in healthcare. *npj Digital Medicine* 2019;**2**:77. doi:10.1038/s41746-019-0155-4

17 Luengo-Oroz M, Hoffmann Pham K, Bullock J, *et al.* Artificial intelligence cooperation to support the global response to COVID-19. *Nature Machine Intelligence* 2020;**2**:295–7. doi:10.1038/s42256-020-0184-3

18 Peiffer-Smadja N, Maatoug R, Lescure F-X, *et al.* Machine Learning for COVID-19 needs global collaboration and data-sharing. *Nature Machine Intelligence* 2020;**2**:293–4. doi:10.1038/s42256-020-0181-6

19 Xie J, Hungerford D, Chen H, *et al.* Development and External Validation of a Prognostic Multivariable Model on Admission for Hospitalized Patients with COVID-19. 2020. doi:10.2139/ssrn.3562456

20 Chow DS, Glavis-Bloom J, Soun JE, *et al.* Development and external validation of a prognostic tool for COVID-19 critical disease. *PLoS One* 2020;**15**:1–11. doi:10.1371/journal.pone.0242953

21 Gupta RK, Harrison EM, Ho A, *et al.* Development and validation of the ISARIC 4C Deterioration model for adults hospitalised with COVID-19: a prospective cohort study. *Lancet Respir Med* 2021;**9**:349–59. doi:10.1016/S2213-2600(20)30559-2

22 Carmichael H, Coquet J, Sun R, *et al.* Learning from past respiratory failure patients to triage COVID-19 patient ventilator needs: A multi-institutional study. *J Biomed Inform* 2021;**119**:103802. doi:10.1016/j.jbi.2021.103802

23 Barish M, Bolourani S, Lau LF, *et al.* External validation demonstrates limited clinical utility of the interpretable mortality prediction model for patients with COVID-19. *Nature Machine Intelligence* 2021;**3**:25–7. doi:10.1038/s42256-020-00254-2

24 Eriksson CO, Stoner RC, Eden KB, *et al.* The Association Between Hospital Capacity Strain and Inpatient Outcomes in Highly Developed Countries: A Systematic Review. *J Gen Intern Med* 2017;**32**:686–96. doi:10.1007/s11606-016-3936-3

25 Emanuel EJ, Persad G, Upshur R, *et al.* Fair Allocation of Scarce Medical Resources in the Time of Covid-19. *N Engl J Med* 2020;**382**:2049–55. doi:10.1056/NEJMsb2005114

26 Vergano M, Bertolini G, Giannini A, *et al.* Clinical ethics recommendations for the allocation of intensive care treatments in exceptional, resource-limited circumstances: the Italian perspective during the COVID-19 epidemic. *Crit Care* 2020;**24**:165. doi:10.1186/s13054-020-02891-w

27 Carenzo L, Costantini E, Greco M, *et al.* Hospital surge capacity in a tertiary emergency referral centre during the COVID-19 outbreak in Italy. *Anaesthesia* 2020;**75**:928–34. doi:10.1111/anae.15072

28 COVIDSurg Collaborative. Elective surgery cancellations due to the COVID-19 pandemic: global predictive modelling to inform surgical recovery plans. *Br J Surg* 2020;**107**:1440–9.

doi:10.1002/bjs.11746

29 Kaufman S, Rosset S, Perlich C, *et al.* Leakage in data mining: Formulation, detection, and avoidance. *ACM Trans Knowl Discov Data* 2012;**6**:1–21. doi:10.1145/2382577.2382579

30 de Mendonça A, Vincent J-L, Suter PM, *et al.* Acute renal failure in the ICU: risk factors and outcome evaluated by the SOFA score. *Intensive Care Med* 2000;**26**:915–21. doi:10.1007/s001340051281

31 Breiman L. Random Forests. *Mach Learn* 2001;**45**:5–32. doi:10.1023/A:1010933404324

32 Hooker G, Mentch L. Please Stop Permuting Features: An Explanation and Alternatives. arXiv. 2019.http://arxiv.org/abs/1905.03151

33 Ferri FJ, Pudil P, Hatef M, *et al.* Comparative study of techniques for large-scale feature selection. In: Gelsema ES, Kanal LS, eds. *Machine Intelligence and Pattern Recognition*. North-Holland 1994. 403–13. doi:10.1016/B978-0-444-81892-8.50040-7

34 Tang S, Davarmanesh P, Song Y, *et al.* Democratizing EHR analyses with FIDDLE: a flexible data-driven preprocessing pipeline for structured clinical data. *J Am Med Inform Assoc* 2020;**27**:1921–34. doi:10.1093/jamia/ocaa139

35 Naeini MP, Cooper GF, Hauskrecht M. Obtaining Well Calibrated Probabilities Using Bayesian Binning. In: *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*. Austin, Texas: : AAAI Press 2015. 2901–2907.

36 Huang Y, Li W, Macheret F, *et al.* A tutorial on calibration measurements and calibration models for clinical prediction models. *J Am Med Inform Assoc* 2020;**27**:621–33. doi:10.1093/jamia/ocz228

37 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**:1129–37. doi:10.1513/AnnalsATS.202006-698OC

38 Python Software Foundation. *Python*. https://www.python.org/

39 Harris CR, Millman KJ, van der Walt SJ, *et al.* Array programming with NumPy. *Nature* 2020;**585**:357–62. doi:10.1038/s41586-020-2649-2

40 Reback J, jbrockmendel, McKinney W, *et al. pandas-dev/pandas: Pandas 1.3.2*. Zenodo 2021. doi:10.5281/ZENODO.3509134

41 McKinney W. Data Structures for Statistical Computing in Python. In: *Proceedings of the 9th Python in Science Conference*. SciPy 2010. doi:10.25080/majora-92bf1922-00a

42 Pedregosa F, Varoquaux G, Gramfort A, *et al.* Scikit-learn: Machine Learning in Python. *J Mach Learn Res* 2011;**12**:2825–30.

43 Rajkomar A, Dean J, Kohane I. Machine Learning in Medicine. *N Engl J Med* 2019;**380**:1347–58. doi:10.1056/NEJMra1814259

44 Angus DC. Randomized Clinical Trials of Artificial Intelligence. JAMA. 2020;**323**:1043–5. doi:10.1001/jama.2020.1039

45 Escobar GJ, Liu VX, Schuler A, *et al.* Automated Identification of Adults at Risk for In-Hospital Clinical Deterioration. *N Engl J Med* 2020;**383**:1951–60. doi:10.1056/NEJMsa2001090

46 Wong A, Otles E, Donnelly JP, *et al.* External Validation of a Widely Implemented Proprietary Sepsis Prediction Model in Hospitalized Patients. *JAMA Intern Med* 2021;**181**:1065–70. doi:10.1001/jamainternmed.2021.2626

47 Davis SE, Lasko TA, Chen G, *et al.* Calibration drift among regression and machine learning models for hospital mortality. In: *AMIA Annual Symposium Proceedings*. American Medical Informatics Association 2017. 625.

48 Van Calster B, Wynants L, Timmerman D, *et al.* Predictive analytics in health care: how can we know it works? *J Am Med Inform Assoc* 2019;**26**:1651–4. doi:10.1093/jamia/ocz130

49 Obermeyer Z, Powers B, Vogeli C, *et al.* Dissecting racial bias in an algorithm used to manage the health of populations. *Science* 2019;**366**:447–53. doi:10.1126/science.aax2342

50 Buolamwini J, Gebru T. Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. In: Friedler SA, Wilson C, eds. *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*. New York, NY, USA: : PMLR 2018. 77–91.http://proceedings.mlr.press/v81/buolamwini18a.html

51 Ashana DC, Anesi GL, Liu VX, *et al.* Equitably Allocating Resources during Crises: Racial Differences in Mortality Prediction Models. *Am J Respir Crit Care Med* 2021;**204**:178–86. doi:10.1164/rccm.202012-4383OC

52 Nguyen NT, Chinn J, Nahmias J, *et al.* Outcomes and Mortality Among Adults Hospitalized With COVID-19 at US Medical Centers. *JAMA Netw Open* 2021;**4**:e210417. doi:10.1001/jamanetworkopen.2021.0417

53 Rosenberg ES, Dufort EM, Udo T, *et al.* Association of Treatment With Hydroxychloroquine or Azithromycin With In-Hospital Mortality in Patients With COVID-19 in New York State. *JAMA* 2020;**323**:2493–502. doi:10.1001/jama.2020.8630

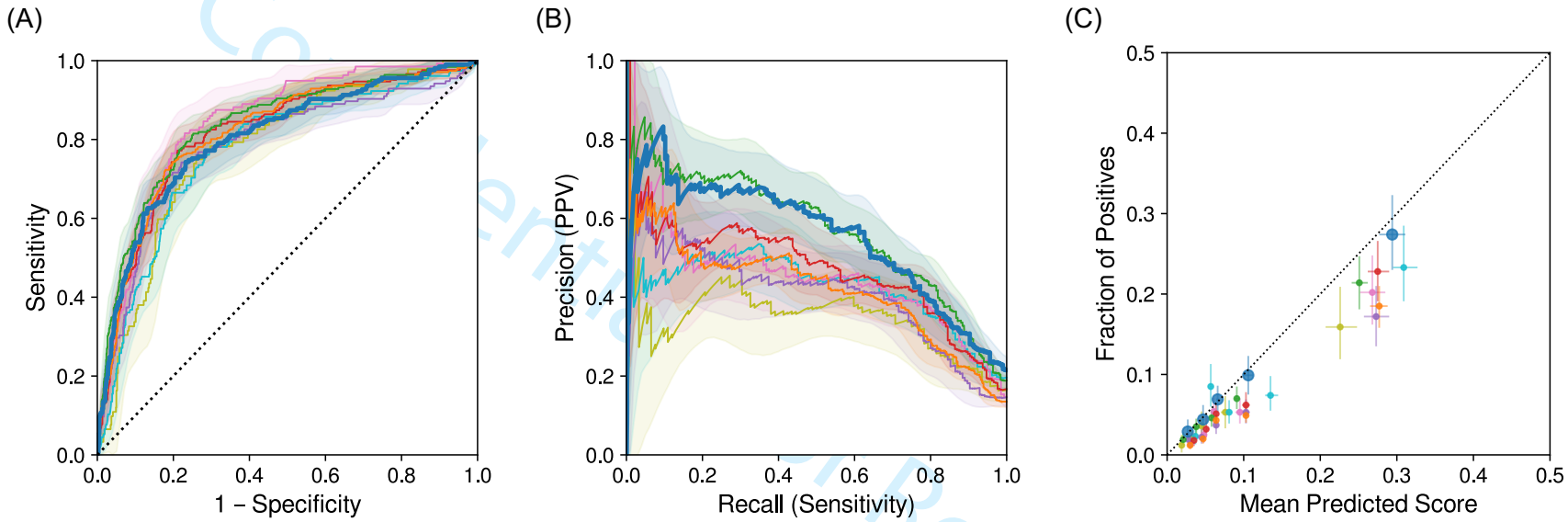
54 Wang M, Zhang J, Ye D, *et al.* Time-dependent changes in the clinical characteristics and prognosis of hospitalized COVID-19 patients in Wuhan, China: A retrospective study. *Clin Chim Acta* 2020;**510**:220–7. doi:10.1016/j.cca.2020.06.051

55 Angeli F, Bachetti T, Maugeri Study Group. Temporal changes in co-morbidities and mortality in patients hospitalized for COVID-19 in Italy. Eur. J. Intern. Med. 2020;**82**:123–5. doi:10.1016/j.ejim.2020.10.019

56 Kip KE, Snyder G, Yealy DM, *et al.* Temporal changes in clinical practice with COVID-19 hospitalized patients: Potential explanations for better in-hospital outcomes. bioRxiv. 2020. doi:10.1101/2020.09.29.20203802

57 Sands KE, Wenzel RP, McLean LE, *et al.* Changes in hospitalized coronavirus disease 2019 (COVID-19) patient characteristics and resource use in a system of community hospitals in the United States. *Infect Control Hosp Epidemiol* 2021;**42**:228–9. doi:10.1017/ice.2020.1264

Figure 1. Model performance across the internal and external validation cohorts. We measure discriminative performance in (A) ROC curves and (B) PR curves. Model calibration is shown in (C) Reliability plots based on quintiles of predicted scores. Results with 95% confidence intervals are summarized in (D). The internal validation cohort at Michigan Medicine (MM) is bolded, while the external validation cohorts A-G are shown in different colors. Overall, discriminative performance and calibration performance was good across institutions. The AUPR varied most in part due to variation in outcome rates.



(D)

		AUROC	AUPR	ECE
●	**MM**	0.804 (0.770, 0.841)	0.549 (0.483, 0.631)	0.007 (0.003, 0.021)
●	**A**	0.816 (0.789, 0.841)	0.418 (0.363, 0.479)	0.042 (0.035, 0.051)
●	**B**	0.834 (0.803, 0.862)	0.567 (0.504, 0.638)	0.015 (0.006, 0.026)
●	**C**	0.816 (0.783, 0.848)	0.467 (0.399, 0.537)	0.027 (0.015, 0.039)
●	**D**	0.785 (0.742, 0.827)	0.405 (0.334, 0.489)	0.042 (0.030, 0.054)
●	**E**	0.843 (0.806, 0.874)	0.451 (0.368, 0.542)	0.031 (0.020, 0.043)
●	**F**	0.774 (0.728, 0.816)	0.419 (0.349, 0.503)	0.041 (0.028, 0.053)
●	**G**	0.777 (0.727, 0.829)	0.338 (0.262, 0.430)	0.024 (0.010, 0.038)

Figure 2. Model discriminative performance (AUROC and AUPR scores) over the year broken down by quarter. The table denotes the legend and the number of hospitalizations included within each cohort in each quarter along with the percentage that met the outcome (in parentheses). The discriminative performance varied most in the second quarter during which there were the fewest number of patients who met the primary outcome. The AUROC across institutions varied little by the fourth quarter or third wave of the pandemic.

		Mar '20 - May '20	Jun '20 - Aug '20	Sep '20 - Nov '20	Dec '20 - Feb '21
●	MM	246 (27.2)	53 (18.9)	287 (18.8)	370 (20.3)
●	A	968 (17.7)	152 (7.9)	282 (12.1)	918 (10.2)
●	B	69 (26.1)	244 (17.2)	337 (16.9)	670 (19.7)
●	C	544 (18.9)	82 (11.0)	146 (13.0)	484 (15.5)
●	D	380 (19.7)	76 (14.5)	141 (12.1)	476 (10.9)
●	E	296 (19.3)	51 (17.6)	140 (6.4)	478 (12.8)
●	F	350 (23.1)	54 (13.0)	93 (21.5)	297 (15.8)
●	G	56 (19.6)	125 (19.2)	122 (14.8)	304 (12.8)

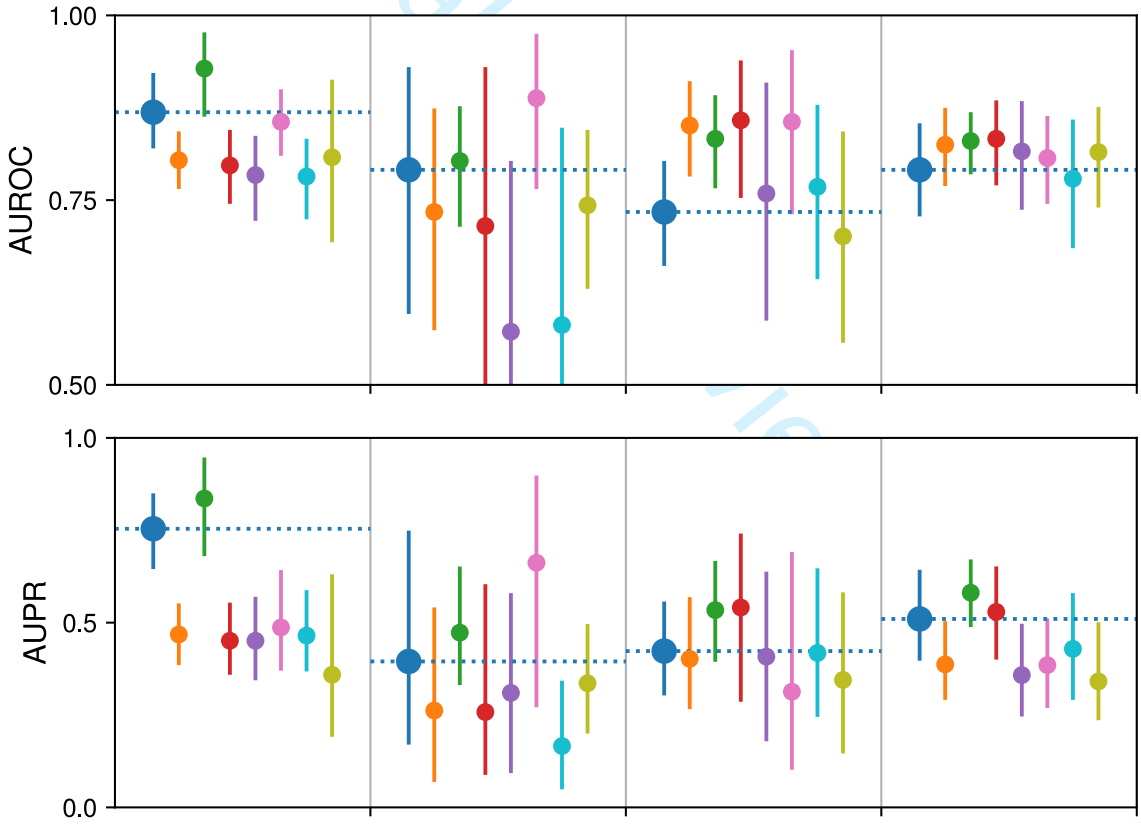
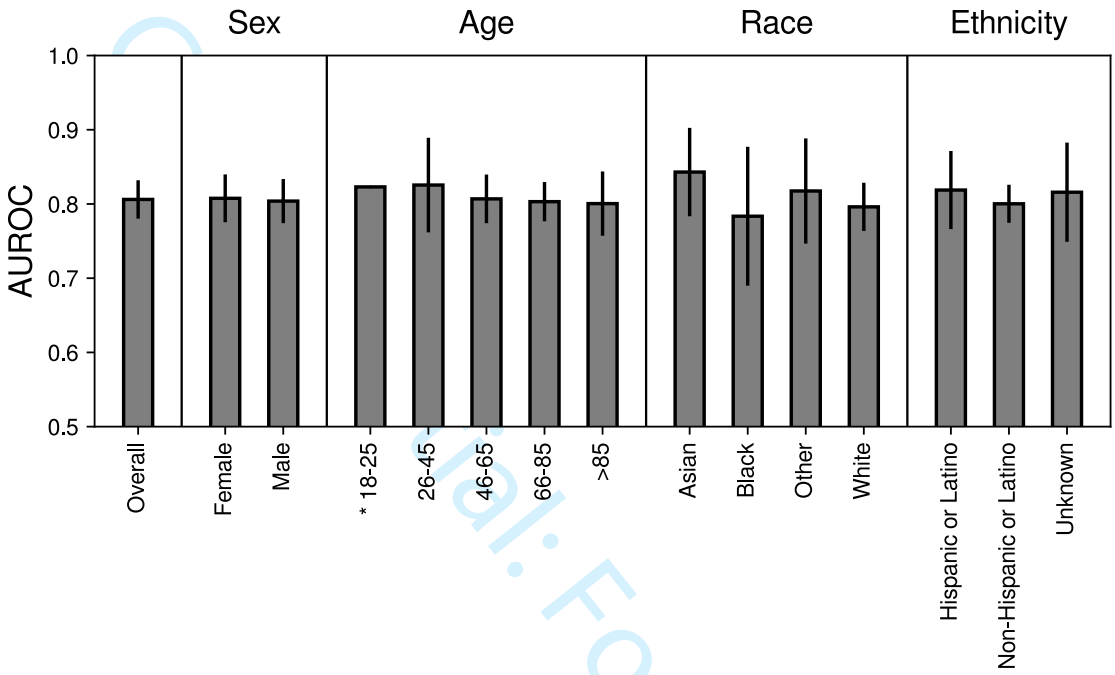
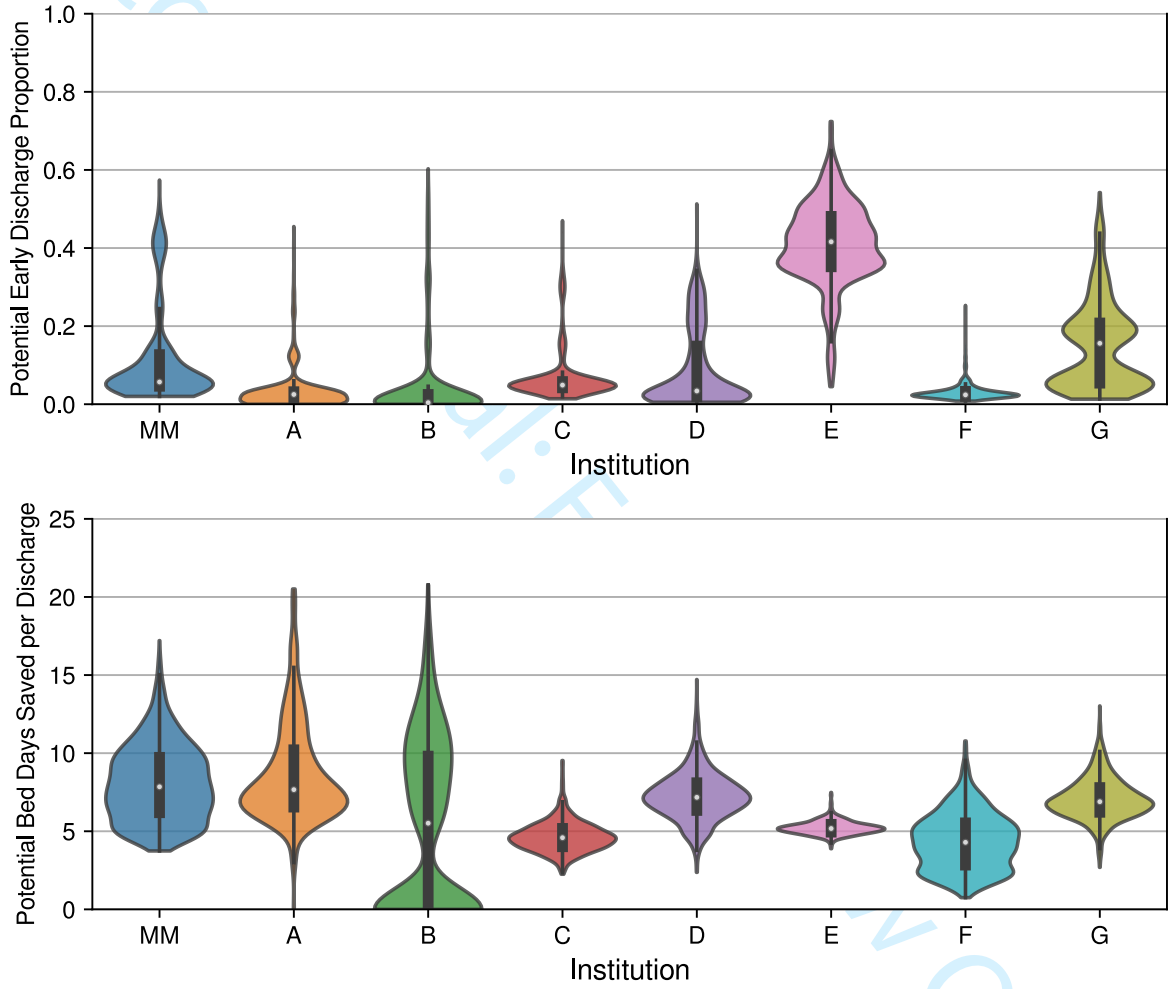


Figure 3. Model discriminative performance (AUROC scores) evaluated across demographic subgroups. Values are macro-average performance across institutions (error bars are ± one standard deviation). Across subgroups the AUROC did not vary significantly from the overall performance.



* No error bar is shown for the 18-25 subgroup because only a single institution had enough positive cases in this subgroup to calculate the AUROC score.

Figure 4. The model can be used to identify potential candidates for early discharge after 48 hours of observation. Using a decision threshold that achieves a negative predictive value of greater or equal to 95%, both the proportion of patients that could be discharged early (top) and the bedtime savings (in days), normalized by the number of correctly discharged hospitalizations at each institution (bottom), are depicted. Results are computed over 1000 bootstrap replications.



Supplement

Supplemental Online Content for:
Kamran F, Tang S, et al. “Early Identification of Hospitalized Patients with COVID-19 at Risk of Clinical Deterioration - A Multi-Site Study”.

Table of Contents

eMethods 1. Details of Internal and External Validation Cohorts.

Inclusion Criteria.

- Michigan Medicine (MM)
  - **COVID Diagnosis**: To identify COVID-19, we included hospitalizations with either (i) a positive laboratory test or (ii) a recorded ICD-10 code for COVID-19 and the absence of a negative laboratory test.
  - **Respiratory Distress:** Adult inpatient hospitalizations in which the patient required supplemental oxygen.
- University of California, San Francisco (UCSF)
  - **COVID diagnosis:** Either “Detected” or “Indeterminate” covid test result, or when patient flagged as having covid from infection control status table.
  - **Respiratory Distress:** Any patient that had value for O2 device (that was not room air) OR (O2 flow rate >0) OR (FiO2 > 21)
- University of Texas, Southwestern (UTSW)
  - **COVID diagnosis:** We included all COVID-19 infections associated with hospital encounters as retrieved from the COVID_19_HSP_INFECTIONS table. Patients are accessible in the table as part of the COVID-19 Hospital Infections registry where patients are added if they have an active or presumed COVID-19 infection flag during the admission.
  - **Respiratory Distress:** Includes all patients requiring supplemental oxygen during admission identified by flowsheet documentation of any oxygen device other than “room air”, any ventilator settings, any O2 flow >0, or O2 concentration >21%.
- Mass General Brigham (MGB)
  - **COVID diagnosis:** We included hospitalizations where the patient had an active COVID-19 or CoV Presumed infection flag at some point during the admission. At MGB, COVID-19 infection flags are automatically added after a positive COVID-PCR test or positive BinaxNOW now antigen assay. CoV-Presumed is applied in the following scenarios 1) symptoms and positive serological assay for SARS-CoV-2, 2) positive antigen assay when symptoms are documented, excluding the BinaxNOW assay; 3) PCR resulting as inconclusive, presumptive positive or NEG late signal (reported only at one institution on the Cepheid GeneXpert assay); 4) positive PCR or BinaxNOW assay in an individual who is between 91-180 days after initial diagnosis of COVID-19 or 5) at the discretion of Infection Control.
  - **Respiratory Distress:** Adult inpatient hospitalizations in which the patient required supplemental oxygen. Supplemental oxygen was defined as having flowsheet documentation of an oxygen device other than “None (Room Air)”.

Outcome Definition. Outcome labels were implemented by each institution individually, as hospitalization-level data were not shared across sites. An initial definition was developed by MM and then adapted by each individual institution in order to ensure that the same outcomes were captured accurately given differences in care processes and informatics infrastructure across institutions. Specific implementation details are summarized below. In general, MV and HHFNC are defined based on clinical events recorded in flowsheets; vasopressor are defined using keyword searches over medication administration records. While the MV and vasopressor definitions are mostly consistent, the HHFNC definition is not identical at each institution due to differences in workflows, though they all correspond to an elevated level of care. Specifically, at some institutions, we have an additional criterion of O2 flow rate ≥ 15L, because at these institutions, nasal cannula with low O2 flow rates were used on the floor but are recorded in the same way as nasal cannula with higher flow rates that are used in the ICU.

- MM
  - **IV vasopressors:** Vasopressors are defined by medication administration records (MAR); we performed a keyword search on the drug name of the MAR for the following well-recognized vasopressors: 'norepinephrine' (aka 'levophed'), 'epinephrine', 'dopamine', 'vasopressin', 'phenylephrine' (aka 'neo-synephrine', 'neosynephrine'), 'angiotensin', and further filtered administrations with route of 'IV' and notgiven = False.
  - **Mechanical Ventilation:** any of the following flowsheet event:
    - "UM IP R CMV START / STOP [Invasive Ventilation Start / Stop]" (313141) with value of "Start"
    - "UM IP R VENT MODE [Vent Mode]" (315640) with a few specific values
    - "UM ED R OXYGEN DEVICE [O2 Device]" 307923 with value 'Ventilator - Emergency Department' or 'Mechanical Ventilation - UH/CVC'
  - **HHFNC:** recorded flowsheet event of "UM ED R OXYGEN DEVICE [O2 Device]" (307923) with value 'Nasal Cannula - Heated High Flow'
- UCSF
  - **IV vasopressors:** Med admin route as (Intravenous, or continuous infusion, or continuous IV infusion, or central venous line induction) and following medications: Dobutamine, Dopamine, Ephedrine, Epinephrine, Milrinone, Norepinephrine, Phenylephrine, Vasopressin.
  - **Mechanical Ventilation:** After excluding patients who had MV on admission (included string "present_on_admission" in values related to intubation), first time where there was a value for "R RT VENT MODE" that was not null
  - **HHFNC:** Either Nasal Cannula or HFNC values for oxygen delivery device with flow rates > 15

- UTSW
  - **IV vasopressors:** Includes all MAR administration of the pressors below based on medication order ID only if route is “intravenous”, medication was given (i.e., excludes the following MAR actions: 'Paused','Stopped','Canceled Entry','Held','HELD BY PROVIDER','Missed'), and rate is >0.
    - '261095', '732983', '249321', '272111' --vasopressin
    - '240032', '240493', '12588', '732981' --norepinephrine
    - '3398', '250088', '244667', '3400', '266933', '735102', '250565', '250964', '732978' --epinephrine
    - '118907','232425','232428','19051' --dobutamine
    - '31759', '231514' --milrinone
    - '232499', '232498', '232500' --dopamine
    - '240509', '7429', '246371', '732982', '240041','734102' --phenylephrine
    - '233968', '230498', '3382' --ephedrine
  - **Mechanical Ventilation:** Includes flowsheet documentation of ventilator mode ('UTSW R ED VENTILATOR MODE') or a ventilator FiO2 ('UTSW R ED VENTILATOR FIO2 (%)')
  - **HHFNC**: Includes flowsheet documentation of an oxygen device of “high-flow nasal cannula” with an O2 flow rate cutoff > 15.
- MGB
  - **IV vasopressors**: Defined as a documented MAR administration of a vasopressor with an associated MAR action indicating that the medication was given (i.e., excluding actions such as “missed” and “held”). Restricted to MAR actions with a documented route of “Intravenous” and a non-zero dose. Vasopressors were defined using pharmaceutical subclasses of Cardiovascular Sympathomimetic - Beta-Adrenergic Agonists, Antidiuretic and Vasopressor Hormones, Cardiovascular Sympathomimetics, and Renin-Angiotensin-Aldosterone System (RAAS) Hormones
  - **Mechanical Ventilation:** Defined as flowsheet documentation of a ventilator mode of ‘AC/VC’, ‘AC/PC’, ‘SIMV/PC’, ‘ASV’, ‘AC/PRVC’, ‘PC-PSV’, ‘AC/VG’, ‘SIMV/PRVC’, ‘SIMV/VC’, or ‘HFJV’
  - **HHFNC**: Defined as flowsheet documentation of an oxygen device of “High Flow Nasal Cannula” or “High flow face mask”. No additional O2 rate cutoffs were used.

eMethods 2. Additional Details on Model Development and Validation.

Variable Selection. First, clinicians reviewed the potential list of EHR variables and removed those which may potentially leak the outcome or themselves are model scores, such as SOFA scores. From here, a model was trained on the remaining EHR variables using the development cohort. EHR variables were sorted based on their permutation importance as measured on the development set.[31,32,59] Variables were added to a set of features one by one, based on their permutation importance, and the model was retrained using just the subset of features. Each retrained model was then evaluated on a small subset of COVID-19 patients from Michigan Medicine (which were subsequently removed from Michigan Medicine's internal validation cohort). Variables were added until the performance on the subset of COVID-19 patients did not increase, resulting in 9 total variables.

External Validation Details. Master Table description

Demographic Subgroups. We considered the following demographic subgroups.

- Age groups are defined by pre-specified bins: 18-25, 26-45, 46-65, 66-85, >85
- Sex: Female, Male
- Race: Asian, Black, White, Other (which includes: American Indian or Alaskan, Native Hawaiian or Other Pacific Islander, Other, Unknown, Patient Refused, More than 1).
- Ethnicity: Hispanic or Latino, Non–Hispanic or Latino, Unknown

Model Training. The goal of the primary prediction task was to identify high-risk patients who deteriorate quickly. Thus, we labeled a hospitalization based on whether or not the patient experienced the composite outcome within five days of hospitalization. We used all 4-hour windows from the time of the first vital sign up until (but not including) either i) 5 days after the first vital sign was measured or ii) the window in which the individual experienced the outcome or was discharged (whichever comes first). We randomly sampled one window per individual hospitalization to include in the training set, such that no individual was represented more than any other. We repeated this process and created 500 different training sets, leading to an ensemble of 500 regularized logistic regression models, whose outputs were averaged to create a final prediction. The model hyperparameter (L2 regularization strength) was selected using 5-fold cross-validation on the first model and applied to the remaining models in the ensemble.

Primary Use-Case Hospitalization-Level Evaluation. To evaluate on a hospitalization level, we swept the decision threshold and identified individuals who exceeded that threshold prior to the endpoint (when outcome is met or when the 5-day mark is reached) as high risk and low risk otherwise. This approach has been used in past work and avoids biasing our evaluations to patient encounters with more windows [Henry et al. 2015, Oh et al. 2018, Singh et al. 2020]. Additionally, at inference time, to ensure the model is not biased by incomplete data, we removed all windows in which a complete 4-hour window of data was unavailable.

Secondary Use Case Evaluation. To evaluate models for the secondary use-case, (i.e., triaging low-risk patients), we consider a situation in which a triaging decision is made 48 hours after the patient's first vital sign is measured. Accordingly, we excluded patient hospitalizations that were no longer eligible for potential triaging at 48 hours (those who met the composite outcome or were discharged within 48h of the patient's first vital sign measurement). For each hospitalization, we make the triaging decision based on the average model prediction within the first 48 hours (excluding incomplete windows). A hospitalization's risk score is defined as their average model score of each complete window within the first 48 hours. To measure the number of hospitalizations we can correctly triage to lower acuity care, we calculated the maximum percentage of hospitalizations correctly flagged as low risk (i.e., those with the lowest average predicted score) where the negative predictive value (NPV) is greater than or equal to 0.95 (i.e., of the hospitalizations flagged as low risk, at least 95% will not meet the outcome during the hospital stay). Moreover, for these hospitalizations, we calculated the potential number of days saved, normalized by the total number of correct discharges, if the flagged individuals were discharged from the hospital at 48 hours into their stay. We repeated the procedure on 1,000 bootstrapped samples of each hospital's cohort and visualized the distributions of potential discharge proportions and potential bed days savings and reported the median values from the bootstrapped results.

Confidence Intervals. For all results, 95% confidence intervals (CIs) were generated using 1,000 bootstrapped samples of each cohort.

eText. Additional Results and Discussion

- A well-validated outcome definition is crucial to external validation. If incorrectly coded, it does not matter how good the model is, evaluation metrics will suffer. While in-hospital mortality is easy to measure, our composite outcome which represents ICU-level care is encoded using proxies such as MV, HHFNC, and IV vasopressors. How these data are recorded in the EHR differed across hospitals.
- Based on existing and new connections formed between different institutions and the relevant access to data each institution has, we identified the sites at which we can rapidly perform the external validation. We first provided a specification document to each institution that describes a unified format containing all information needed to perform the evaluation. Researchers at each institution performed their own cohort data extraction from EHR databases and outcome definitions and collated everything into a unified format. Model parameters for each of the 500 models along with the necessary code (including a standard feature processing procedure) were packaged into a transferable computer program by MM, which was sent to each institution. Researchers at each institution then ran the program on their own infrastructure and transferred back only model results; no identifiable information was shared. This procedure was done quickly (within a month) and involved less risk of PHI-related issues compared to sharing raw patient data (which involves signing data use agreements with multiple institutions).

eTable 1. Characteristics of the development cohort and comparison with the internal validation cohort. Both cohorts are from Michigan Medicine. Statistically significant differences (at α=0.001 with a Bonferroni correction for multiple hypotheses) are denoted by *.

Institution	Development	Internal Validation	p-value
Number of patients	24,419	887	-
Number of hospitalizations	35,040	956	-
Median age in years [IQR]	63 [51-74]	64 [52–75]	-
Age Group (%)			0.02
[18, 25]	1,275 (3.6)	17 (1.8)	
(25, 45]	5,114 (14.6)	129 (13.5)	
(45, 65]	13,060 (37.3)	374 (39.1)	
(65, 85]	13,064 (37.3)	365 (38.2)	
>85	2,432 (6.9)	70 (7.3)	
Sex (%)			0.01
Female	16,877 (48.2)	420 (43.9)	
Male	18,163 (51.8)	536 (56.1)	
Race (%)			<0.0001*
White	29,402 (83.9)	649 (67.9)	
Black	3,954 (11.3)	187 (19.6)	
Asian	625 (1.8)	30 (3.1)	
Other/Unknown	1,059 (3.0)	90 (9.4)	
Ethnicity (%)			-
Hispanic or Latino	-	34 (3.6)	
Not Hispanic or Latino	-	883 (92.4)	
Other/Unknown	-	39 (4.1)	
Median LOS in hours [IQR]	97 [55–173]	138 [83–261]	-
Outcome ever (%)			<0.0001*
Death	963 (2.7)	60 (6.3)	
MV	2,341 (6.7)	98 (10.3)	
IV Vaso	1,320 (3.8)	87 (9.1)	
HHFNC	1,858 (5.3)	218 (22.4)	
Primary Outcome <= 5 days	3,757 (10.7)	206 (21.6)	<0.0001*
Reason for composite outcome (% of outcomes)			<0.0001*
Death	252 (6.7)	5 (2.4)	
MV	1,737 (46.2)	20 (9.7)	
IV Vaso	454 (12.1)	9 (4.4)	
HHFNC	1,314 (35.0)	172 (83.5)	

Acronyms: IQR, interquartile range; LOS, Length-of-Stay; MV, Mechanical Ventilation; IV, Intravenous Vasopressors, HHFNC, Heated High-Flow Nasal Cannula.

eTable 2. P-values for pairwise comparisons of characteristics between the internal validation cohort and each external validation cohort. We applied chi-square tests for homogeneity to compare categorical demographic variables. Every external validation cohort differed in at least one demographic dimension. Statistically significant differences (at α=0.001 with a Bonferroni correction for multiple hypotheses) are denoted by *.

Characteristic	MM vs A	MM vs B	MM vs C	MM vs D	MM vs E	MM vs F	MM vs G
Sex	0.6	0.3	0.6	0.01	0.3	0.5	0.003
Age Group	0.02	0.009	3e-6*	0.09	5e-11	3e-21*	2e-5*
Race	2e-43*	3e-10*	1e-9*	4e-11*	2e-7	2e-16*	2e-70*
Ethnicity	2e-51*	1e-52*	2e-49*	4e-28*	1e-15	1e-14*	1e-45*
Has Outcome (Ever)	7e-36*	0.05	4e-14*	8e-14*	4e-8*	2e-12*	0.03
Has Primary Outcome	6e-9*	0.1	0.002	3e-5*	2e-5*	0.3	0.002

Acronyms: MM, Michigan Medicine.

eTable 3. P-values for pairwise comparisons of the reasons for meeting the composite outcome, between the internal validation cohort and the development cohort, as well as between the internal cohort and each external validation cohort. We applied chi-square tests for homogeneity to compare the reason for outcome. Statistically significant differences (at α=0.006 with a Bonferroni correction for multiple hypotheses) are denoted by *.

Reason for Outcome	MM vs DEV	MM vs A	MM vs B	MM vs C	MM vs D	MM vs E	MM vs F	MM vs G
p-value	2e-41*	3e-30*	0.7	1e-11*	2e-15*	4e-7*	9e-12*	0.007

Acronyms: MM, Michigan Medicine; DEV, Development cohort.

eTable 4. Estimated 95% confidence intervals of the performance difference between the internal validation cohort and each external validation cohort. The difference is significant if the interval does not overlap with zero (denoted by *).

Institution	MM vs A	MM vs B	MM vs C	MM vs D	MM vs E	MM vs F	MM vs G
Difference in AUROC	[-0.05, 0.03]	[-0.08, 0.02]	[-0.06, 0.04]	[-0.04, 0.08]	[-0.08, 0.01]	[-0.02, 0.09]	[-0.03, 0.09]
Difference in AUPR	[0.04, 0.23] *	[-0.13, 0.08]	[-0.01, 0.19]	[0.04, 0.25] *	[-0.01, 0.21]	[0.03, 0.24] *	[0.10, 0.32] *

Acronyms: MM, Michigan Medicine; AUROC, Area Under the Receiver Operating Characteristic; AUPR: Area Under the Precision Recall Curve.

eTable 5. Estimated 95% confidence intervals (99.8% CIs with Bonferroni correction) of the performance difference during a specific time period relative to overall performance, within each validation cohort. No difference is statistically significant (the intervals all overlap with zero).

Institution	MM	A	B	C	D	E	F	G
Mar '20 – May '20	[-0.03, 0.16]	[-0.08, 0.06]	[-0.01, 0.18]	[-0.10, 0.05]	[-0.12, 0.11]	[-0.06, 0.10]	[-0.10, 0.10]	[-0.13, 0.22]
Jun '20 – Aug '20	[-0.34, 0.20]	[-0.36, 0.14]	[-0.19, 0.08]	[-0.55, 0.18]	[-0.60, 0.20]	[-0.22, 0.16]	[-0.71, 0.15]	[-0.23, 0.13]
Sept '20 – Nov '20	[-0.18, 0.04]	[-0.08, 0.13]	[-0.12, 0.10]	[-0.11, 0.17]	[-0.39, 0.18]	[-0.20, 0.19]	[-0.19, 0.19]	[-0.30, 0.13]
Dec '20 – Feb '21	[-0.13, 0.11]	[-0.10, 0.09]	[-0.08, 0.07]	[-0.09, 0.11]	[-0.09, 0.14]	[-0.14, 0.07]	[-0.15, 0.16]	[-0.09, 0.16]

Acronyms: MM, Michigan Medicine.

eTable 6. Estimated 95% confidence intervals (99.8% CIs with Bonferroni correction) of the performance difference of each demographic subgroup relative to overall performance, within each validation cohort. No subgroup is significantly different from overall performance in terms of AUROC.

Institution	MM	A	B	C	D	E	F	G
Sex:F	[-0.14, 0.11]	[-0.07, 0.10]	[-0.13, 0.06]	[-0.10, 0.10]	[-0.08, 0.14]	[-0.10, 0.12]	[-0.13, 0.12]	[-0.19, 0.10]
Sex:M	[-0.09, 0.08]	[-0.09, 0.07]	[-0.05, 0.10]	[-0.10, 0.07]	[-0.14, 0.08]	[-0.09, 0.08]	[-0.12, 0.11]	[-0.12, 0.13]
Age:17-25	N/A	[-0.61, 0.22]	N/A	N/A	N/A	N/A	N/A	N/A
Age:25-45	[-0.25, 0.11]	[-0.03, 0.14]	[-0.16, 0.12]	[-0.19, 0.15]	[-0.13, 0.21]	[-0.28, 0.17]	[-0.16, 0.23]	[-0.44, 0.26]
Age:45-65	[-0.15, 0.08]	[-0.13, 0.06]	[-0.05, 0.11]	[-0.09, 0.13]	[-0.20, 0.13]	[-0.14, 0.10]	[-0.15, 0.15]	[-0.16, 0.14]
Age:65-85	[-0.07, 0.13]	[-0.07, 0.09]	[-0.12, 0.06]	[-0.11, 0.08]	[-0.18, 0.11]	[-0.11, 0.11]	[-0.12, 0.13]	[-0.17, 0.15]
Age:85-1000	[-0.13, 0.20]	[-0.19, 0.10]	[-0.49, 0.18]	[-0.21, 0.14]	[-0.27, 0.20]	[-0.14, 0.13]	[-0.27, 0.10]	[-0.22, 0.18]
Race:Asian	[-0.58, 0.24]	[-0.16, 0.16]	[-0.63, 0.20]	[-0.19, 0.15]	[-0.01, 0.27]	N/A	[-0.10, 0.28]	[-0.16, 0.16]
Race:Black	[-0.09, 0.15]	[-0.12, 0.16]	[-0.12, 0.11]	[-0.23, 0.14]	[-0.09, 0.21]	[-0.27, 0.14]	[-0.41, 0.21]	[-0.63, 0.17]
Race:Other	[-0.32, 0.20]	[-0.10, 0.09]	[-0.10, 0.12]	[-0.32, 0.15]	[-0.09, 0.18]	[-0.03, 0.16]	[-0.33, 0.08]	[-0.14, 0.15]
Race:White	[-0.12, 0.07]	[-0.09, 0.07]	[-0.09, 0.08]	[-0.06, 0.10]	[-0.19, 0.07]	[-0.09, 0.07]	[-0.07, 0.12]	[-0.18, 0.15]
Ethnicity:Hispanic	[-0.60, 0.23]	[-0.06, 0.10]	[-0.11, 0.09]	[-0.08, 0.12]	[-0.12, 0.16]	[-0.02, 0.18]	[-0.29, 0.16]	[-0.12, 0.18]
Ethnicity:Non-Hispanic	[-0.08, 0.08]	[-0.06, 0.06]	[-0.06, 0.08]	[-0.10, 0.08]	[-0.14, 0.08]	[-0.11, 0.07]	[-0.10, 0.10]	[-0.14, 0.12]
Ethnicity:Unknown	[-0.39, 0.24]	[-0.20, 0.11]	[-0.39, 0.20]	[-0.46, 0.22]	N/A	[-0.10, 0.20]	[-0.40, 0.20]	N/A

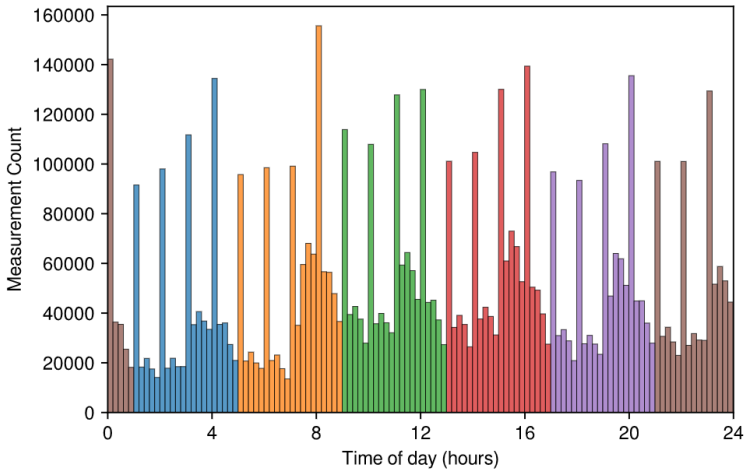
Acronyms: MM, Michigan Medicine; F, Female; M, Male; AUROC, Area Under the Receiver Operating Characteristic.

eTable 7. Estimated 95% confidence intervals (99.8% CIs with Bonferroni correction) of the performance difference between White and each other race subgroup, within each validation cohort. The difference is significant if the interval does not overlap with zero (denoted by *).

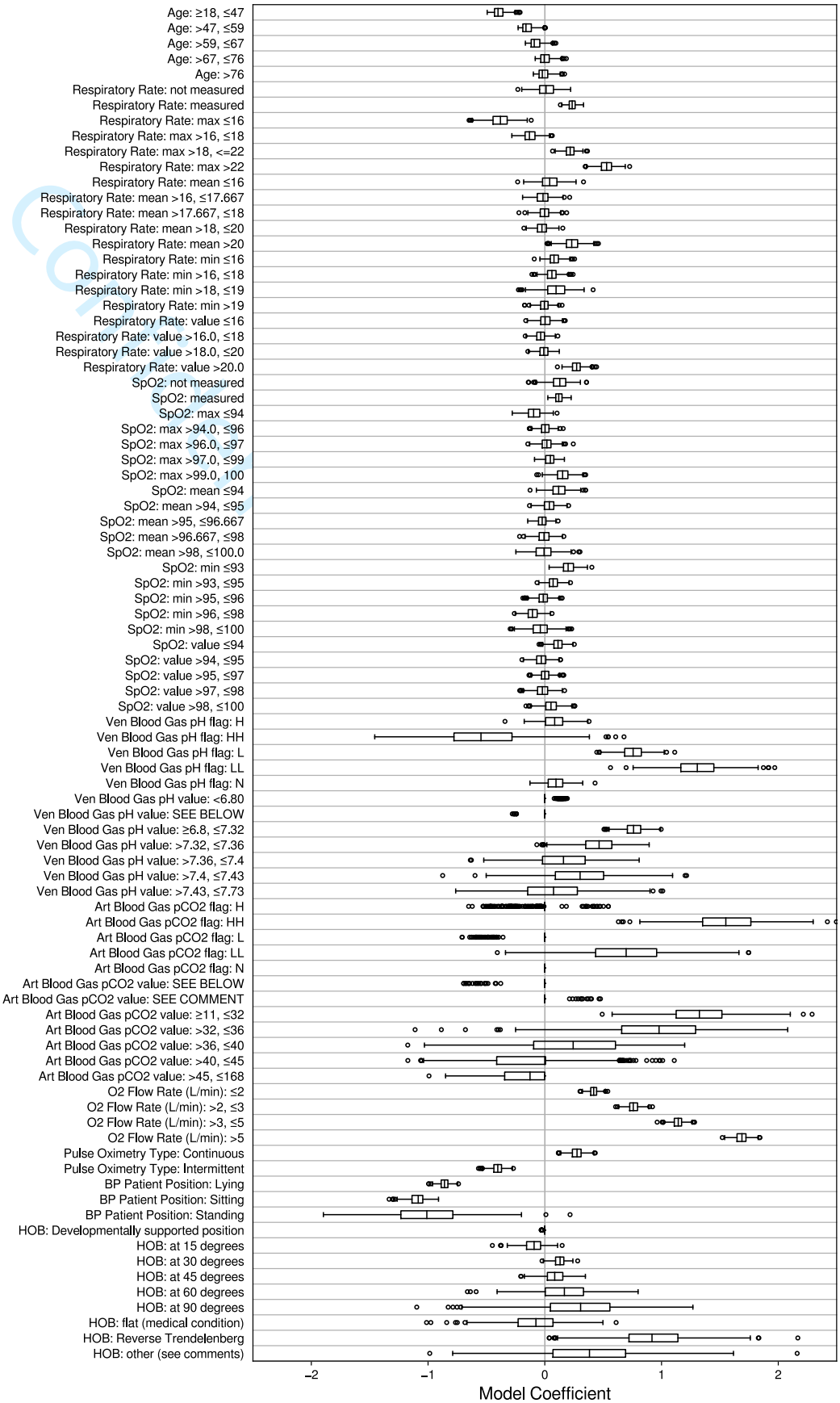
Institution	MM	A	B	C	D	E	F	G
W-A	[-0.26, 0.53]	[-0.19, 0.13]	[-0.18, 0.45]	[-0.17, 0.19]	[-0.36, -0.04] *	N/A	[-0.25, 0.08]	[-0.21, 0.20]
W-B	[-0.17, 0.06]	[-0.18, 0.11]	[-0.10, 0.12]	[-0.12, 0.22]	[-0.26, 0.03]	[-0.13, 0.23]	[-0.17, 0.36]	[-0.17, 0.66]
W-O	[-0.20, 0.30]	[-0.12, 0.07]	[-0.13, 0.08]	[-0.14, 0.22]	[-0.28, 0.00]	[-0.17, 0.02]	[-0.02, 0.39]	[-0.22, 0.15]

Acronyms: MM, Michigan Medicine; A, Asian; B, Black; O, Other races; W, White.

eFigure 1. Measurement frequency of patient heart rate throughout different times of the day, in the development cohort (Michigan Medicine, 2015-2019). Based on the empirical measurement frequency of important vital signs, we defined 4-hour time windows with respect to time points of a day at 1am, 5am, 9am, 1pm, 5pm, and 9pm. These time points correspond to right after the measurement "peaks" and were selected with the feasibility of real-time deployment of the system in mind.

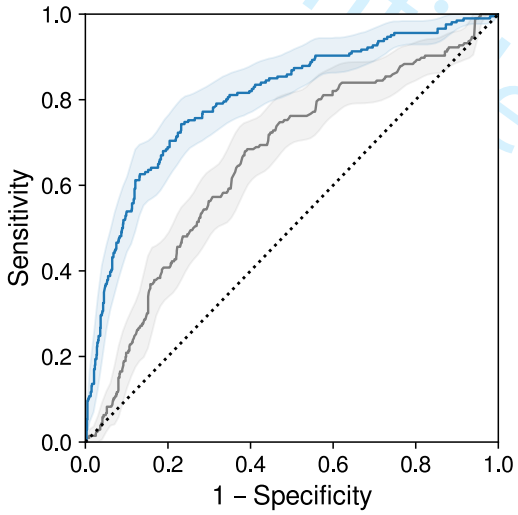


eFigure 2. Weights of the 88 features over 500 regularized logistic regression models in the ensemble.

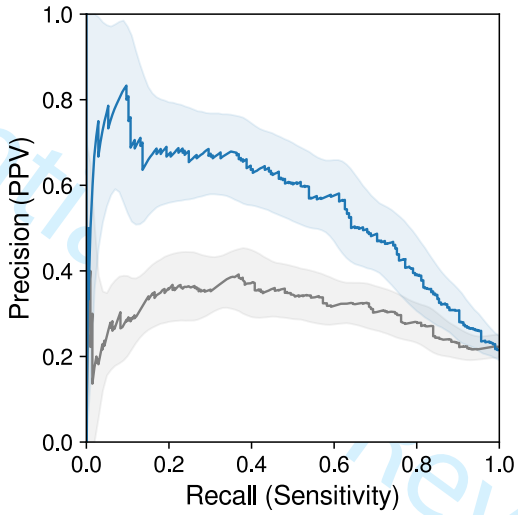


eFigure 3. Model performance of MCURES model (shown in blue) and Epic Deterioration Index (shown in gray) on the MM internal validation cohort. We measure discriminative performance in (A) ROC curves and (B) PR curves. Model calibration is shown in (C) Reliability plots based on quintiles of predicted scores. Legend and results with 95% confidence intervals are summarized in (D). The MCURES model outperforms the Epic Deterioration Index in terms of both discriminative performance and calibration performance.

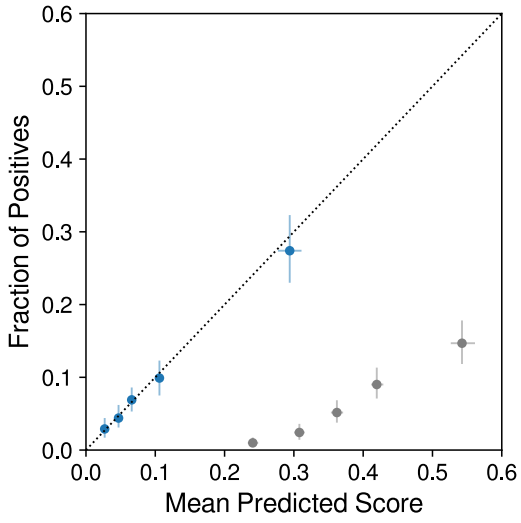
(A)



(B)



(C)



(D)

		AUROC	AUPR	ECE
●	**MCURES**	0.804 (0.770, 0.841)	0.549 (0.483, 0.631)	0.007 (0.003, 0.021)
●	**EDI**	0.657 (0.618, 0.701)	0.309 (0.264, 0.362)	0.310 (0.297, 0.322)

EDI, Epic Deterioration Index.