JMIR Publications
The Leading eHealth Publisher

Development and validation of simplified machine learning algorithms to predict prognosis of hospitalized COVID-19 patients: a multi-center, retrospective study

Fang He, John H Page, Kerry R Weinberg, Anirban Mishra

Submitted to: Journal of Medical Internet Research
on: June 24, 2021

Preprint
JMIR Publications

Disclaimer: © The authors. All rights reserved. This is a privileged document currently under peer-review/community review. Authors have provided JMIR Publications with an exclusive license to publish this preprint on it's website for review purposes only. While the final peer-reviewed paper may be licensed under a CC BY license on publication, at this stage authors and publisher expressively prohibit redistribution of this draft paper other than for review purposes.

Table of Contents

Preprint JMIR Publications

Development and validation of simplified machine learning algorithms to predict prognosis of hospitalized COVID-19 patients: a multi-center, retrospective study

Fang He[1] BChE, PhD; John H Page[2] MD, SCD; Kerry R Weinberg[3] BSc, MBA, MSci; Anirban Mishra[3] BSc, MSc

[1]Amgen Inc Center for Observational Research | Digital Health & Innovation South San Francisco US
[2]Amgen Inc Center for Observational Research Thousand Oaks US
[3]Amgen Inc Digital Health & Innovation Thousand Oaks US

Corresponding Author:
Fang He BChE, PhD
Amgen Inc
Center for Observational Research | Digital Health & Innovation
1120 Veterans Blvd
South San Francisco
US

Preprint JMIR Publications

Abstract

Background: The current COVID-19 pandemic is unprecedented; under resource-constrained setting, predictive algorithms can help to stratify disease severity, alerting physicians of high-risk patients, however there are few risk scores derived from a substantially large EHR dataset, using simplified predictors as input.

Objective: To develop and validate simplified machine learning algorithms which predicts COVID-19 adverse outcomes, to evaluate the AUC (area under the receiver operating characteristic curve), sensitivity, specificity and calibration of the algorithms, to derive clinically meaningful thresholds.

Methods: We conducted machine learning model development and validation via cohort study using multi-center, patient-level, longitudinal electronic health records (EHR) from Optum® COVID-19 database which provides anonymized, longitudinal EHR from across US. The models were developed based on clinical characteristics to predict 28-day in-hospital mortality, ICU admission, respiratory failure, mechanical ventilator usages at inpatient setting. Data from patients who were admitted prior to Sep 7, 2020, is randomly sampled into development, test and validation datasets; data collected from Sep 7, 2020 through Nov 15, 2020 was reserved as prospective validation dataset.

Results: Of 3.7M patients in the analysis, a total of 585,867 patients were diagnosed or tested positive for SARS-CoV-2; and 50,703 adult patients were hospitalized with COVID-19 between Feb 1 and Nov 15, 2020. Among the study cohort (N=50,703), there were 6,204 deaths, 9,564 ICU admissions, 6,478 mechanically ventilated or EMCO patients and 25,169 patients developed ARDS or respiratory failure within 28 days since hospital admission. The algorithms demonstrated high accuracy (AUC = 0.89 (0.89 - 0.89) on validation dataset (N=10,752)), consistent prediction through the second wave of pandemic from September to November (AUC = 0.85 (0.85 - 0.86) on post-development validation (N= 14,863)), great clinical relevance and utility. Besides, a comprehensive 386 input covariates from baseline and at admission was included in the analysis; the end-to-end pipeline automates feature selection and model development process, producing 10 key predictors as input such as age, blood urea nitrogen, oxygen saturation, which are both commonly measured and concordant with recognized risk factors for COVID-19.

Conclusions: The systematic approach and rigorous validations demonstrate consistent model performance to predict even beyond the time period of data collection, with satisfactory discriminatory power and great clinical utility. Overall, the study offers an accurate, validated and reliable prediction model based on only ten clinical features as a prognostic tool to stratifying COVID-19 patients into intermediate, high and very high-risk groups. This simple predictive tool could be shared with a wider healthcare community, to enable service as an early warning system to alert physicians of possible high-risk patients, or as a resource triaging tool to optimize healthcare resources. Clinical Trial: N/A

(JMIR Preprints 24/06/2021:31549)
DOI: https://doi.org/10.2196/preprints.31549

Preprint Settings

1) Would you like to publish your submitted manuscript as preprint?

✓ **Please make my preprint PDF available to anyone at any time (recommended).**
Please make my preprint PDF available only to logged-in users; I understand that my title and abstract will remain visible to all users.
Only make the preprint title and abstract visible.
No, I do not wish to publish my submitted manuscript as a preprint.

2) If accepted for publication in a JMIR journal, would you like the PDF to be visible to the public?

✓ **Yes, please make my accepted manuscript PDF available to anyone at any time (Recommended).**
Yes, but please make my accepted manuscript PDF available only to logged-in users; I understand that the title and abstract will remain v
Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in <a href="http

Preprint JMIR Publications

Original Manuscript

Preprint
JMIR Publications

Original Paper

Development and validation of simplified machine learning algorithms to predict prognosis of hospitalized COVID-19 patients

Authors

Authors

Fang He, PhD[1,2*]; John H. Page, MD[1*]; Kerry Weinberg MS[2]; Anirban Mishra, MS[2].

Affiliations

[1]Center for Observational Research, Amgen, Inc., Thousand Oaks, CA; [2]Digital Health & Innovation, Amgen Inc., Thousand Oaks, CA.

Corresponding Author:

Fang He

fhe01@amgen.com

650-447-6132

1120 Veterans Blvd

South San Francisco, CA 94080

John H. Page

jopage@amgen.com

805-490-5527

1 Amgen Center Drive, B38-4B

Thousand Oaks, CA 91320

Preprint JMIR Publications

ABSTRACT

Background

The current COVID-19 pandemic is unprecedented; under resource-constrained setting, predictive algorithms can help to stratify disease severity, alerting physicians of high-risk patients, however there are few risk scores derived from a substantially large EHR dataset, using simplified predictors as input.

Objective

To develop and validate simplified machine learning algorithms which predicts COVID-19 adverse outcomes, to evaluate the AUC (area under the receiver operating characteristic curve), sensitivity, specificity and calibration of the algorithms, to derive clinically meaningful thresholds.

Methods

We conducted machine learning model development and validation via cohort study using multi-center, patient-level, longitudinal electronic health records (EHR) from Optum® COVID-19 database which provides anonymized, longitudinal EHR from across US. The models were developed based on clinical characteristics to predict 28-day in-hospital mortality, ICU admission, respiratory failure, mechanical ventilator usages at inpatient setting. Data from patients who were admitted prior to Sep 7, 2020, is randomly sampled into development, test and validation datasets; data collected from Sep 7, 2020 through Nov 15, 2020 was reserved as prospective validation dataset.

Results

Of 3.7M patients in the analysis, a total of 585,867 patients were diagnosed or tested positive for SARS-CoV-2; and 50,703 adult patients were hospitalized with COVID-19 between Feb 1 and Nov 15, 2020. Among the study cohort (N=50,703), there were 6,204 deaths, 9,564 ICU admissions, 6,478 mechanically ventilated or EMCO patients and 25,169 patients developed ARDS or respiratory failure within 28 days since hospital admission. The algorithms demonstrated high accuracy (AUC =

0.89 (0.89 - 0.89) on validation dataset (N=10,752)), consistent prediction through the second wave of pandemic from September to November (AUC = 0.85 (0.85 - 0.86) on post-development validation (N= 14,863)), great clinical relevance and utility. Besides, a comprehensive 386 input covariates from baseline and at admission was included in the analysis; the end-to-end pipeline automates feature selection and model development process, producing 10 key predictors as input such as age, blood urea nitrogen, oxygen saturation, which are both commonly measured and concordant with recognized risk factors for COVID-19.

Conclusions

The systematic approach and rigorous validations demonstrate consistent model performance to predict even beyond the time period of data collection, with satisfactory discriminatory power and great clinical utility. Overall, the study offers an accurate, validated and reliable prediction model based on only ten clinical features as a prognostic tool to stratifying COVID-19 patients into intermediate, high and very high-risk groups. This simple predictive tool could be shared with a wider healthcare community, to enable service as an early warning system to alert physicians of possible high-risk patients, or as a resource triaging tool to optimize healthcare resources.

Trial Registration

N.A.

Key words

COVID-19; predictive algorithm; prognostic model; machine learning

Word count

441 words

Preprint JMIR Publications

INTRODUCTION

The COVID-19 pandemic has impacted more than 200 countries, claimed more than 3 million lives, presenting an urgent threat to global health. Under resource constrained settings, a validated model using large-scale real-world data (RWD) to predict COVID-19 prognosis can rapidly identify the individuals who are at risk of COVID-19 adverse outcomes and mortality, therefore they could benefit from early interventions.

Several studies have derived prognostic predictors for COVID-19, however currently there are few COVID-19 risk-calculation tools with simplified predictors for stratification which leverages on a substantially large US EHR dataset of statistically meaningful size[1,2]. The Acute Physiology and Chronic Health Evaluation (APACHE) II score[3] has been widely used to predict in-hospital mortality, and has been found to predict mortality in COVID-19 patients, out-performing SOFA[4] and CURB-65[5] scores in a retrospective study of 154 patients in China[6]. COVID-GRAM[2] is a web-based calculator to estimate the occurrence of ICU admission, mechanical ventilation, or death in hospitalized patients with COVID-19; it has been validated in a study of nearly 1,600 patients in China. The 4C (Coronavirus Clinical Characterization Consortium) Mortality Score[1] developed by ISARIC WHO CCP-UK study is a risk stratification tool to predict in-hospital mortality by categorizing patients at low, intermediate, high, or very high risk of death. Separately an accurate, machine learning based COVID-19 mortality prediction model has been developed based on data from Mount Sinai Health System, however the validation dataset is limited in size[7].

The objective of this paper is to develop and validate simplified and parsimonious predictive algorithms, leveraging large size, near real-time RWD as risk stratification methodology to identify patients, who are at heightened risk of (1) mortality; (2) ICU admission; (3) composite of invasive mechanical ventilation/ECMO; (4) composite of ARDS/respiratory failure, which can be easily

integrated into hospital EMR system as a risk stratification and triaging tool.

METHODS

Data source

This is a retrospective observational cohort analysis of multi-center, longitudinal, anonymized patient-level data from Optum® Electronic Health Record (EHR) COVID-19 database. It includes demographics, insurance status, medication prescription, vital signs, coded diagnoses, procedures, lab results, visits, encounters and providers. Currently there are 3,702,050 patients in the data release dated Jan 27, 2021. Given de-identified data is used for the study, it was exempt of IRB approval.

Study period

The study period is from Feb 1, 2020 to Jan 27, 2021. A baseline of up to 1 year prior to and including index date was used for assessment of demographic, lifestyle factors, and comorbidity at baseline. Subjects were followed up to 28 days from admission, unless they were censored by in-hospital mortality or discharged.

Participants

Study cohort consists of hospitalized COVID-19 patients aged 18 and older, with a confirmed diagnosis or positive test of COVID-19 infection. A COVID-19 diagnosis was defined as the first occurrence on or after Feb 1, 2020 of any of the following: 1) positive result from SARS-CoV-2 viral RNA or antigen tests; 2) ICD-10-CM diagnosis code U07.1, J12.81, J12.89, or J20; 3) ICD-10-CM code B97.29 or B34.20 occurring on or before April 30, 2020. Patients were admitted to hospital no earlier than 10 days prior to and no later than 28 days after COVID-19 diagnosis.

Patients were excluded for any of the following: 1) missing age or sex; 2) under the age of 18; 3)

diagnosis or procedure codes for labor and delivery during hospitalization; 4) diagnosis codes for trauma, injury, fracture or poisoning during the first two days of hospitalization; 5) less than 10 weeks of follow-up between their first COVID-19 diagnosis date or hospital admission date, and the last database refresh date (Jan 27, 2021) (Figure 1). Additional sensitivity analysis was conducted between final study cohort (N=50,703) and patients who tested positive for SARS-CoV-2 (N=38,277), a subset of the former.

Index date

The index date is defined as hospital admission date.

Sample size

The initial anonymized data for 3,702,050 patients from 885,677 providers and 2,465 delivery networks for the study period Feb 1, 2020 – Jan 27, 2021 was transferred from Optum®; among which, 585,867 patients were diagnosed or tested positive for SARS-CoV-2 infection.

In the final cohort that satisfied the study criteria (N=50,703), data from patients with an index date prior to Sep 7, 2020 was referred to as model development dataset (N=35,840), which was randomly sampled without replacement using 28-day in-hospital mortality as stratification factor, into 40% training dataset (N=14,336), 30% test dataset (N=10,752) for hyperparameter tuning and threshold calculation, and 30% validation dataset (N=10,752). The sampling ratio is determined such that the validation or test dataset alone can satisfy the sample size requirement; the minimum sample size is estimated to be 8,605, assuming a pre-determined sensitivity of 0.7 and the prevalence of all-cause mortality of 15% with 95% confidence interval and maximum marginal error of estimate of 2.5%[8].

Furthermore, a second validation dataset consisting of patients with index date from Sep 7 to Nov 15,

2020, was referred to as independent validation dataset (N=14,863) (Figure 1). The demographic and clinical characteristics of patients in the training, test and two validation datasets are summarized in Table 1.

Outcome

The outcomes are 28-day in-hospital (1) all-cause mortality; (2) ICU admission; (3) composite of invasive mechanical ventilation or ECMO; (4) composite of ARDS and respiratory failure. These were assessed as dichotomous outcomes and individually modelled. Outcome-specific exclusions were applied as appropriate to include only incident outcomes.

Covariates

A total of 386 study covariates consisting of patients' baseline demographics (age, sex, CENSUS division, insurance status, race, ethnicity), lifestyle factors (smoking status, BMI), comorbidities (including atrial fibrillation cancer history, cerebrovascular disease, chronic kidney disease stage I-V, COPD, coronary artery disease, Type I/II diabetes mellitus, HIV, stroke, etc.), baseline medication (including antidiabetics, anticoagulants, antihypertensives, antiplatelets, steroids, etc.) within 12 months prior to index date, and vital signs (blood pressures, heart rate, pulse, respiration rate, temperature), laboratory values (including albumin, alanine transaminase, aspartate aminotransferase, total bilirubin, B-type natriuretic peptide, blood urea nitrogen, chloride, creatinine, c reactive protein, d-dimer, fibrinogen, hemoglobin, lymphocyte, monocyte, neutrophil, oxygen saturation platelet count, arterial blood pH, etc.) and treatment (including diuretics, DMARDS, steroids, etc.) administered on the day of hospital admission were included in the analysis. Concretely, baseline medication, comorbidity, and post-admission treatment were expressed as dichotomous variables; categorical variables were converted to dummy variables; numerical variables were used without standardization, unless when fitting to penalized (Lasso or Ridge)

logistic regression models, numerical covariates were normalized using a min-max standardization to speed up convergence.

Missing data

One of the challenges of working with real world data is the missing covariates. Assuming covariates are missing at random, multiple imputation by chained equations via random forest[9] was used to impute covariates with missing values. Ten complete datasets each with ten iterations were imputed with predictive mean matching using available covariates while excluding the outcome variables. The prediction performances of sparsity-aware models (XGB[10]) between imputed and non-imputed dataset were compared in the sensitivity analysis.

Given the intention to develop an algorithm of great relevance to as many patients as possible, we have restricted the model input to covariates with a minimum of 70% coverage in the study cohort. Sensitivity analysis shows the inclusion of additional covariates with higher degrees of missingness (i.e., varying the cutoffs from 10% to 90%) has limited success in further improving model performance (Supplementary Figure 6). Further, this may increase the sensitivity of the models to biases due to non-ignorable missingness data.

Model development

We have applied a systematic approach to model development and validation. A framework of six machine learning algorithms (XGB[10,11], penalized Logistic Regression[12,13] with Lasso[14] or Ridge loss[11], Random Forest[11,15,16], Decision Tree[17], lightGBM[18]) have been adopted to develop interpretable models to predict the prognosis of COVID-19.

In the preliminary analysis, the most performant algorithm was selected from the candidate

algorithms; prior to model training, hyperparameter optimization via grid search, ranging from 96 folds to 243 folds was performed on six candidate algorithms individually for full and simplified models. The full model uses all the available 386 input features after extraction and transformation, while simplified model recursively eliminates aforementioned input to yield a maximum of 20 variables[19]. The algorithm with best performance (area under the receiver operating characteristic (AUC), Brier score[20] and calibration[21]) on validation datasets was selected for the final analysis.

In the final analysis, model input is further iteratively reduced to a maximum of 5 variables with a step size of 1; 100 individual runs were performed at each step, with retuned model parameters every 5 steps. The selected features were pooled and plotted in frequency heatmap with corresponding AUC.

The model performance is evaluated against outcome variables in validation datasets via AUC, Brier score and calibration curve. 95% confidence intervals for AUC and Brier score are calculated based on percentiles from bootstrapped resampling with replacement (bootstrap sample size = 2,000) without bias correction or acceleration[22]. The calibration curves (number of discretized bins = 10) were plotted for all the runs.

Model validation

Rigorous validation analysis was performed to ensure robustness and reliability of the predictions. Both full and simplified models of six candidate algorithms were trained and validated during model development phase with data from Feb 1 to Sep 6, where the validation dataset was held out from model training and used solely for reporting the performance. Furthermore, model has been additionally validated externally, using patients' data collected from Sep 7 to Nov 15, 2020 to demonstrate consistent model performance through the subsequent wave of pandemic. Model

discrimination was performed on imputed validation datasets by assessing AUC on the stratified analysis by sex, age and racial groups.

Model benchmark

The performance of the risk prediction models has been benchmarked to 1) the baseline model; 2) published COVID-19 prognostic scores. The baseline model was developed using XGB with optimized hyperparameters on age and sex only. Evaluation metrics including AUC, sensitivity, specificity and decision curve analysis were assessed to compare the performance and utility of prognostic scores (APACHE II[3,23], CURB-65[5,24,25], E-CURB[26], NEWS2 score[25,27–29], ROX index[29,30], 4C mortality score[1,25]). AUC is reported based on complete case data from both validation datasets, and no imputation was performed.

Predictors

Feature importance is ranked by their Shapley values[31] from validation datasets in SHAP summary plot. Shapley value calculates fair contribution and the extent of predictors towards the model output[32]. It measures feature importance by the magnitude and the direction of contributions. The dependence between model prediction and age is plotted with age on x-axis, its impact on prediction represented by Shapley value on y-axis for every patient, colored by the magnitude of a second feature (blood urea nitrogen (BUN), respiration rate, pulse, lymphocyte count) individually.

ROC curve analysis

We adopted two approaches in determining the optimal threshold cut-off on ROC curve. Assuming the sensitivity and specificity were weighted equally without ethical, cost and prevalence constraints, the optimal cut-off is at the location where Youden’s index (sum of specificity and sensitivity - 1) is maximized at the test dataset[33–36]. This approach relies solely on the predictive accuracy of a

model, and consequences of the predictions (*i.e.* cost of false positives and false negatives) are not considered. In the second approach, clinical utility-based decision theory was used in developing cost-sensitive cut-offs, where it builds in disease prevalence, and costs of false positive and false negatives of specific diagnostic scenario[33,37].

Decision curve analysis assists in clinical judgment and comparison about the relative value of benefits associated with use of a clinical prediction tool[38,39]. The standardized net benefit of full model, simplified model (with ten input variables) and selected benchmark prognostic scores were calculated and plotted across probabilities. The benchmark models which use point scores were calibrated to validation data prior to decision curve analysis.

RESULTS

Patient characteristics

Figure 1 shows patient attrition flowchart, and the workflow of model development and validation is in Figure 2. Patients' baseline and clinical characteristics at admission are summarized in Table 1. Test and validation datasets are largely homogeneous to the training dataset; however, the second independent validation dataset which was collected later in pandemic from September to November, presents more differences in geographic locations (a decline in proportion of patients in Middle Atlantic from 22.7% to 8.7% after Sep 7, and an increase in West North Central from 9.8% to 24.6%) and racial distribution (proportion of Caucasian increased from 53.9% to 72.9%). However, the overall mortality and remains consistent. Hypertension (58.4%), obesity (47.2%), diabetes (33.9%), chronic kidney disease (19.9%) and coronary artery disease (17.9%) are the common comorbidities among the cohort.

Table 1. Demographic and clinical characteristics of hospitalized COVID patients at baseline and admission.

	Training dataset (N=14,336)	Test dataset (N=10,752)	Validation dataset (N=10,752)	Post-development validation dataset (N=14,863)
Age at baseline				
Mean, year	60.9 ± 17.2	60.9 ± 17.2	60.8 ± 17.1	63.8 ± 16.8
Distribution, n (%)				
18-34	1,231 (8.6%)	920 (8.6%)	911 (8.5%)	1,015 (6.8%)
35-49	2,383 (16.6%)	1,780 (16.6%)	1,840 (17.1%)	1,893 (12.7%)
50-64	4,337 (30.3%)	3,193 (29.7%)	3,293 (30.6%)	4,110 (27.7%)
65-74	2,922 (20.4%)	2,296 (21.4%)	2,141 (19.9%)	3,325 (22.4%)
75-84	2,165 (15.1%)	1,589 (14.8%)	1,606 (14.9%)	2,943 (19.8%)
85+	1,298 (9.1%)	974 (9.1%)	961 (8.9%)	1,577 (10.6%)
Sex at baseline, n (%)				
Male	7,473 (52.1%)	5,619 (52.3%)	5,629 (52.4%)	7,645 (51.4%)
Female	6,863 (47.9%)	5,133 (47.7%)	5,123 (47.6%)	7,218 (48.6%)
Race at baseline, n (%)				
African American	3,466 (24.2%)	2,669 (24.8%)	2,668 (24.8%)	1,867 (12.6%)
Asian	368 (2.6%)	268 (2.5%)	276 (2.6%)	216 (1.5%)
Caucasian	7,779 (54.3%)	5,734 (53.3%)	5,795 (53.9%)	10,832 (72.9%)
Other/Unknown	2,723 (19.0%)	2,081 (19.4%)	2,013 (18.7%)	1,948 (13.1%)
Census Division at baseline, n (%)				
East North Central	3,778 (26.4%)	2,942 (27.4%)	2,908 (27.0%)	4,174 (28.1%)
East South Central	1,010 (7.0%)	708 (6.6%)	754 (7.0%)	1,205 (8.1%)
Middle Atlantic	3,221 (22.5%)	2,488 (23.1%)	2,423 (22.5%)	1,290 (8.7%)
Mountain	496 (3.5%)	355 (3.3%)	363 (3.4%)	923 (6.2%)
New England	1,042 (7.3%)	705 (6.6%)	763 (7.1%)	769 (5.2%)
Pacific	475 (3.3%)	331 (3.1%)	317 (2.9%)	345 (2.3%)
South Atl/West South Crl	2,454 (17.1%)	1,802 (16.8%)	1,810 (16.8%)	2,120 (14.3%)
West North Central	1,396 (9.7%)	1,067 (9.9%)	1,060 (9.9%)	3,594 (24.2%)
Other/Unknown	464 (3.2%)	354 (3.3%)	354 (3.3%)	443 (3.0%)
BMI at baseline				
Mean, kg/m^2	31.0 ± 8.5	30.9 ± 8.3	31.2 ± 8.6	31.6 ± 8.7
Distribution, n (%)				
Underweight	352 (2.5%)	235 (2.2%)	221 (2.1%)	304 (2.0%)
Healthy weight	2,526 (17.6%)	1,873 (17.4%)	1,833 (17.0%)	2,283 (15.4%)
Overweight	3,697 (25.8%)	2,878 (26.8%)	2,838 (26.4%)	3,679 (24.8%)
Obese	3,041 (21.2%)	2,247 (20.9%)	2,344 (21.8%)	3,228 (21.7%)
Morbidly Obese	3,679 (25.7%)	2,739 (25.5%)	2,742 (25.5%)	4,069 (27.4%)
Unknown	1,041 (7.3%)	780 (7.3%)	774 (7.2%)	1,300 (8.7%)
Comorbidity at baseline[a], n (%)				
Cerebrovascular disease	676 (4.7%)	502 (4.7%)	501 (4.7%)	894 (6.0%)
Chronic kidney disease	2808 (19.6%)	2058 (19.1%)	2040 (19.0%)	3127 (21.0%)
Congestive heart failure	2137 (14.9%)	1534 (14.3%)	1553 (14.4%)	2369 (15.9%)
Coronary artery disease	2430 (17.0%)	1797 (16.7%)	1800 (16.7%)	2969 (20.0%)
Diabetes mellitus	4831 (33.7%)	3636 (33.8%)	3586 (33.4%)	5408 (34.5%)
Hypertension	8173 (57.0%)	6091 (56.6%)	6063 (56.4%)	8852 (59.6%)
Solid tumor	830 (5.8%)	606 (5.6%)	619 (5.8%)	1052 (7.1%)
Transplant history	28 (0.2%)	16 (0.1%)	20 (0.2%)	12 (0.1%)
28-day Outcomes, n (%)				
All-cause mortality	1,769 (12.3%)	1,326 (12.3%)	1,327 (12.3%)	1,782 (12.0%)
ICU admission	2,813 (19.6%)	2,181 (20.3%)	2,148 (20.0%)	2,422 (16.3%)
ARDS | respiratory failure	7,276 (50.8%)	5,500 (51.2%)	5,384 (50.1%)	7,009 (47.2%)
ECMO | mechanical ventilation	1,962 (13.7%)	1,535 (14.3%)	1,498 (13.9%)	1,483 (10.0%)
Vital at admission				
Diastolic blood pressure (mm Hg)[b]	73.0 (56.0, 90.0)	73.0 (56.0, 90.0)	73.0 (56.0, 90.0)	73.0 (56.0, 90.0)
Systolic blood pressure (mm Hg)[b]	125.0 (100.0, 154.0)	125.0 (101.0, 155.0)	125.0 (101.0, 154.0)	128.0 (103.0, 159.0)
Pulse (bpm)[b]	85.0 (64.0, 110.0)	85.0 (64.0, 110.0)	85.0 (64.0, 110.0)	81.0 (61.0, 107.6)
Respiratory rate (breaths/min)[b]	19.0 (16.0, 28.0)	19.0 (16.0, 28.0)	19.0 (16.0, 28.0)	18.0 (16.0, 25.0)
Temperature (°C)[b]	36.8 (36.3, 37.9)	36.8 (36.3, 37.9)	36.8 (36.3, 37.8)	36.7 (36.2, 37.7)
Lab[a] at admission				
Alkaline phosphatase (IU/L)	77.0 (49.0, 137.0)	76.0 (49.0, 136.0)	76.0 (48.0, 135.0)	78.0 (50.0, 134.0)
Alanine aminotransferase (IU/L)	28.0 (12.0, 79.0)	29.0 (12.0, 80.0)	28.0 (12.0, 79.0)	27.0 (12.0, 68.0)
Aspartate aminotransferase (IU/L)	37.0 (18.0, 95.0)	36.0 (18.0, 97.0)	36.0 (18.0, 95.0)	34.0 (18.0, 80.0)
Albumin (g/dL)	3.5 (2.7, 4.2)	3.6 (2.7, 4.2)	3.6 (2.7, 4.2)	3.6 (2.8, 4.2)
Anion gap (mEq/L)	12.0 (7.0, 17.0)	12.0 (7.0, 17.0)	12.0 (7.0, 17.0)	12.0 (7.0, 16.0)
Blood urea nitrogen (mg/dL)	16.0 (8.0, 47.0)	17.0 (8.0, 46.0)	16.0 (8.0, 47.0)	18.0 (9.0, 44.0)

Bicarbonate (mmol/L)	24.0 (19.0, 29.0)	24.0 (19.0, 29.0)	24.0 (19.0, 29.0)	24.0 (19.0, 29.0)
Bilirubin, total (mg/dL)	0.6 (0.3, 1.2)	0.6 (0.3, 1.2)	0.6 (0.3, 1.2)	0.6 (0.3, 1.1)
C reactive protein, CRP (mg/dL)	85.0 (10.3, 229.0)	82.2 (11.0, 218.0)	82.0 (10.2, 220.0)	73.0 (10.0, 206.6)
Chloride (mmol/L)	101.0 (94.0, 108.0)	101.0 (94.0, 108.0)	101.0 (94.0, 108.0)	101.0 (94.0, 107.0)
Glucose (mg/dL)	120.0 (91.0, 242.0)	121.0 (92.0, 236.0)	121.0 (92.0, 240.6)	122.0 (91.2, 244.0)
Hemoglobin (g/dL)	13.2 (10.0, 15.5)	13.2 (10.1, 15.6)	13.2 (10.2, 15.7)	13.2 (10.1, 15.6)
Lymphocyte (%)	14.1 (5.4, 30.0)	14.8 (5.6, 30.7)	14.6 (5.8, 30.2)	14.1 (5.3, 30.0)
Monocyte (%)	7.1 (3.1, 12.9)	7.0 (3.2, 12.6)	7.1 (3.2, 12.7)	7.8 (3.6, 13.1)
Neutrophil (%)	75.8 (57.0, 88.0)	75.0 (57.0, 88.0)	75.2 (57.0, 88.0)	75.0 (57.0, 88.0)
Platelet count ($\times 10^9$/L)	210.0 (125.0, 351.0)	210.0 (127.0, 348.0)	211.0 (126.0, 351.0)	205.0 (124.0, 335.0)
Potassium (mmol/L)	3.9 (3.3, 4.8)	3.9 (3.3, 4.8)	3.9 (3.3, 4.8)	3.9 (3.3, 4.7)
Protein, total (g/dL)	7.2 (6.2, 8.2)	7.2 (6.2, 8.2)	7.3 (6.2, 8.2)	7.1 (6.2, 8.0)
RDW-CV (%)	13.9 (12.4, 17.0)	13.8 (12.4, 16.9)	13.8 (12.4, 17.0)	13.8 (12.4, 16.7)
Sodium (mmol/L)	136.0 (130.0, 141.0)	136.0 (131.0, 142.0)	136.0 (131.0, 141.0)	136.0 (131.0, 141.0)
Oxygen saturation pulse oximeter (%)	96.0 (91.0, 99.0)	96.0 (90.0, 99.0)	96.0 (91.0, 99.0)	95.0 (90.0, 99.0)
Oxygen saturation pulse oximeter[b] (%)	95.0 (87.0, 99.0)	95.0 (87.0, 99.0)	95.0 (87.0, 99.0)	95.0 (87.0, 99.0)
Oxygen saturation pulse oximeter[c] (%)	93.0 (84.0, 97.0)	93.0 (84.0, 97.0)	93.0 (84.0, 97.0)	92.0 (83.0, 97.0)
White blood cell count ($\times 10^9$/L)	7.1 (4.0, 14.1)	7.1 (4.0, 13.9)	7.0 (4.0, 13.8)	6.9 (3.9, 13.5)

[a] *non-exhaustive list*

[b] *first measurement on the day of hospital admission*

[c] *minimum measurement on the day of hospital admission*

The study cohort is defined as adult hospitalized patients with COVID-19 who were either diagnosed with relevant diagnosis codes or tested positive for SARS-CoV-2 viral RNA or antigen tests. In the subgroup analysis of patients who tested positive for SARS-CoV-2, model performances of these two groups (i.e., overall cohort and tested positive subgroup) were largely similar with < 1% difference in AUC across all outcomes for full and simplified models (Supplementary Figure 4).

Algorithm and predictor selection

In the preliminary analysis, all the candidate algorithms perform comparably on two validation datasets, with < 3% difference in AUC for all outcomes between full and simplified models (Supplementary Figure 1). Of the six candidate machine learning algorithms, boosting-based algorithms (XGB[10] and lightGBM[18]) performed consistently better[40] for both full and preliminary simplified models (N=20) with less computation time, produced well calibrated probabilities; XGB was selected given it has been validated in similar approach[7,41,42]. With adequate model calibration and low Brier score, no adjustment or calibration was subsequently

performed.

In the final analysis, 100 individual runs of recursive elimination are pooled at each step between 5- to 20-input model with an increment of 1, and the selected predictors were analyzed in the frequency heatmap (Supplementary Figure 7). The parsimonious model (N10) incorporates the top ten predictors excluding non-modifiable factors such as diagnosis month or CENSUS division. Specifically, the model which predicts all-cause mortality incorporates age, systolic and diastolic blood pressures, respiration rate, pulse, temperature BUN, SpO_2, albumin and presence of any major cognitive disorder (including dementia, Parkinson's disease and Alzheimer's disease). The magnitude and direction of individual feature contribution to prediction are inferred from summary plot of Shapley values sorted by descending order of feature impact (Supplementary Figure 9); an increase in age[43–45] and BUN[45–47] and AST[45,48], a decrease in oxygen saturation[7,49] and platelet count[26,50,51] and albumin[45,52] are associated with increase in mortality risk. The model performances on both validation datasets are summarized in Table 2.

Table 2. Summary of model performances (AUC and Brier Score) on validation dataset 1 (model development phase) and validation dataset 2 (post-development phase) in final analysis.

		AUC (95% CI)		Brier Score (95% CI)	
Outcome	Model	Validation dataset 1	Validation dataset 2	Validation dataset 1	Validation dataset 2
All-cause mortality	Full model	88.7% (88.4%, 89.0%)	85.4% (85.1%, 85.7%)	0.071 (0.070, 0.072)	0.079 (0.078, 0.080)
	N10 model	87.6% (87.2%, 87.9%)	84.3% (84.0%, 84.6%)	0.074 (0.073, 0.075)	0.081 (0.080, 0.081)
ICU admission	Full model	79.7% (79.4%, 80.1%)	77.7% (77.3%, 78.0%)	0.123 (0.122, 0.124)	0.115 (0.114, 0.115)
	N10 model	73.6% (73.2%, 74.0%)	73.5% (73.2%, 73.9%)	0.138 (0.137, 0.139)	0.123 (0.122, 0.124)
Respiratory failure[a]	Full model	82.3% (82.0%, 82.5%)	80.7% (80.5%, 80.9%)	0.172 (0.171, 0.173)	0.180 (0.179, 0.181)
	N10 model	79.5% (79.2%, 79.7%)	78.1% (77.9%, 78.3%)	0.185 (0.184, 0.186)	0.192 (0.191, 0.193)
Mechanical ventilation[b]	Full model	83.6% (83.3%, 84.0%)	81.1% (80.8%, 81.5%)	0.090 (0.089, 0.091)	0.074 (0.074, 0.075)
	N10 model	78.1% (77.7%, 78.5%)	76.6% (76.2%, 77.1%)	0.101 (0.100, 0.101)	0.081 (0.081, 0.082)

[a] *refers to composite of respiratory failure and ARDS*

[b] *refers to composite of invasive mechanical ventilation and ECMO*

DISCUSSION

Predictors

A major strength of this study is the use of near real-time, large-size EHR data, resulting in predictors that are highly representative and relevant to clinical practice. We have restricted the analysis to commonly measured covariates with less than 30% of missing values among the cohort. A higher coverage cutoff precludes key predictors such as oxygen saturation[7,49], respiration rate[53] and BUN[46,47] leading to degradation of model performance (Supplementary Figure 6).

Among 6,478 mechanically ventilated patients, 2,018 (31.2%) were intubated on the first day of admission; similarly, among 9,564 ICU patients, 4,745 (49.6%) patients were admitted to ICU on day one. This indicates patients could be at different phases of their disease trajectories when admitted to hospital, with some being in critical condition. The treatment that these patients received at admission, for instance oxygen support, intubation merely reflects the selective treatment choices given to the more severe patients. Therefore, we have removed post-admission treatment from model input in final analysis; it has minimally impacted the model performance on all-cause mortality.

The parsimonious model consists of clinical features from patient demographics (age), comorbidity (any major cognitive disorder including dementia, Parkinson's disease and Alzheimer's disease), vital signs (systolic and diastolic blood pressures, respiration rate, pulse, temperature) and lab measurements (BUN, SpO_2, albumin), which are already available or commonly measured at clinical settings, and concordant with recognized risk factors for community-acquired pneumonia or COVID-19[1,5,26].

Age is identified as a crucial predictor for adverse outcomes[44], in particular risks of all-cause mortality and ARDS increase almost monotonically with age. However, the relationship between age and other outcomes such as ICU admission, invasive mechanical ventilation/ECMO is more complex, as these outcomes are more closely associated with the availability of healthcare resources such as ventilator and ICU rooms. This is more noticeable for elderly patients above 75 years, who are disadvantaged for mechanical ventilation and ICU though they are at highest mortality risk (Supplementary Figure 8). This is potentially due to the scarcity of healthcare resources during pandemic, yet our models are capable of predicting these outcomes adequately.

Model performance

We have adopted a systematic framework of model development, including a variety of tree-, boosting-, and ensemble-based machine learning models, combined with rigorous validation on statistically meaningful sample size. The model predicts 28-day in-hospital mortality accurately (AUC = 0.88 (95% CI 0.87 – 0.88) on validation dataset) and reliably through the second wave of pandemic (AUC = 0.84 (95% CI 0.84 – 0.85) on independent validation dataset). Given this dataset was acquired later in time from September to November and more likely to suffer from data lag, the completeness and accuracy of outcome data is hypothesized to contribute to the decrease in model performance; a subgroup analysis on patients with the complete clinical features shows an improved performance (AUC = 0.89 (95% CI 0.88 – 0.90) (Table 3).

Table 3. Comparison with existing risk scores evaluated on validation datasets to predict 28-day all-cause mortality. Sensitivity and specificity were evaluated at two different thresholds.

Risk score	AUC (95% CI)	Scenario (Dexamethasone)		Scenario (Youden)		N[a]
		sensitivity	specificity	sensitivity	specificity	
APACHE II	72.3% (69.5%, 74.9%)	66.2%	68.5%	92.4%	26.0%	1,769

ROX	68.5% (67.0%, 70.0%)	28.2%	92.7%	54.2%	78.3%	16,640
CURB-65	78.7% (77.6%, 79.7%)	36.2%	92.4%	77.2%	69.1%	15,001
ECURB	81.9% (80.3%, 83.3%)	63.4%	83.4%	87.3%	61.3%	5,772
NEWS2 Score	82.9% (81.7%, 84.2%)	51.6%	91.2%	75.0%	77.0%	14,112
4C Mortality Score	82.2% (80.7%, 83.5%)	62.3%	83.8%	71.8%	75.7%	6,979
Baseline Model	73.8% (73.2%, 74.5%)	44.8%	83.4%	80.2%	54.9%	25,615
Full Model	89.2% (88.1%, 90.3%)	63.1%	92.2%	85.2%	76.4%	8,493
N10 Model	88.9% (88.0%, 90.0%)	65.9%	90.9%	81.4%	79.3%	10,688

[a] *Number of hospitalized patients in both validation datasets with complete case.*

We also examined discriminatory capacity in subgroups stratified by sex, race, and age group separately. It predicts all-cause mortality similarly among men (AUC=0.84 (95% CI 0.84 – 0.84)) and women (AUC=0.84 (95% CI 0.84 – 0.85)) and is marginally more predictive among Asians (AUC=0.86 (95% CI 0.85 – 0.87)) compared with African Americans (AUC=0.83 (95% CI 0.83 – 0.84)) and Caucasians (AUC=0.84 (95% CI 0.84 – 0.84)). Given age is an important predictor, the model is more sensitive towards elderly cohort (more accurately ruling out negative cases), conversely more specific towards younger cohort (more accurately ruling in the positive cases).

The model shares commonalities (e.g., age, respiration rate, blood pressures, pulse, BUN, SpO_2, albumin) with existing prognostic scores for community acquired pneumonia or COVID-19[5,26,27]; however, with automated feature selection from comprehensive input covariates, and machine learning algorithm, it compares favorably with existing scores across diagnostic statistics (AUC, sensitivity, specificity) (Table 3) and shows greater clinical utility across a wide range of probability thresholds (Figure 5).

Clinical application

When applying the model to clinical setting, threshold selection is of great practical importance in producing dichotomous predictions. In the data-driven, cost-agnostic approach, threshold is derived numerically from AUC curve which maximizes Youden's index[34] (p=0.13). When the model is applied to inform clinical decision making, such as identifying patients for dexamethasone treatment,

insights from relevant clinical trials could guide threshold calculation. For instance, the findings from RECOVERY trial[54], a large-enrollment, randomized controlled trial of dexamethasone, indicates a mortality risk reduction of 4.84% among patients who received oxygen therapy or mechanically ventilated compared to control group; conversely an increase in mortality risk of 3.74% among patients who require no oxygen. When the model is applied clinically as prognostic tool to identify patients who will receive dexamethasone, cost expressed as an increase in mortality risk of false negative (i.e., misclassifying patients as low risk, therefore they missed dexamethasone treatment) is 4.84%, and cost of false positive (i.e., misclassifying patients as high risk) is 3.74%. Given the mortality rate of 24.8% from RECOVERY trial[54], threshold is found from AUC curve at p=0.33 where the slope of curve[33,37] is (0.0374/0.0484)×(1-0.248)/(0.248) = 2.35.

These two thresholds (p=0.13, and p=0.24) are similar to the intermediate and high risk cutoffs used to define the severity of pneumonia[1,55,56]. Based on these approaches, we derived two clinically meaningful thresholds (Table 4), stratifying patients into (1) low-to-intermediate risk ($p \leq 0.13$, observed mortality rate = 3.9%); (2) high risk ($0.13 < p \leq 0.24$, observed mortality rate = 19.2%); and (3) very high risk ($p > 0.24$, observed mortality rate = 51.9%). Scenario-based threshold can be substituted with appropriate clinical trial insights according to different treatment options.

Table 4. Mortality rate comparison across different risk groups from validation and post-development validation cohorts. Three risk groups were defined as (1) low-to-intermediate risk group (probability ≤ 0.13); (2) high risk (0.13 < probability ≤ 0.24), and (3) very high risk (p > 0.24). The threshold probabilities are obtained from ROC analysis which either (1) maximizes Youden index (p=0.13); or (2) defined by clinical utility of dexamethasone (p=0.24) from RECOVERY trial.

Risk group	Validation		Post-development validation	
	Number of patients (%)	Number of deaths (%)	Number of patients (%)	Number of deaths (%)
Low – intermediate	8,065 (75.0%)	315 (3.9%)	11,049 (74.3%)	512 (4.6%)
High	1,170 (10.9%)	225 (19.2%)	1,743 (11.7%)	327 (18.8%)
Very high	1,517 (14.1%)	787 (51.9%)	2,071 (13.9%)	943 (45.5%)

Overall	10,752	1,327	14,863	1,782

Strengths

The strengths of this research include the large size of dataset, longitudinal nature and near real-time update of the data release. The Optum® database provides patient-level information with a diverse mix of geographic regions, insurance types, socioeconomic status, and ethnicity. A comprehensive list of 386 input covariates from baseline and at admission was included in the analysis based on epidemiological and clinical characteristics of COVID-19 cases; the end-to-end pipeline automates feature selection and model development process, producing risk factors which are both commonly measured at admission with wide coverage among study cohort and concordant with similar risk scores. This helps to improve the usability of the model without extensive EMR integration or feeding the model with continuous data streams. The systematic approach and rigorous validations demonstrate consistent model performance to predict even beyond the time period of data collection, with satisfactory discriminatory power and great clinical utility. Overall, the study offers an accurate, validated and reliable prediction model based on only ten clinical features as a prognostic tool to stratifying COVID-19 patients into intermediate, high and very high-risk groups. We envision this model to be used on the day of hospital admission at inpatient setting where resource triaging is most relevant and early identification of high-risk patient is the key.

Limitations

There are several limitations in our study. Firstly, Optum® COVID-19 database being an EHR database, may not capture patients' entire interaction with healthcare systems because patients can switch between different hospitals or healthcare systems. This impacts several aspects of the study, from assessment of baseline comorbidity and co-medication, to capture of outcomes during follow-up. Though we have identified a minimum of ten-week follow-up from database refresh date to COVID-19 diagnosis date to allow for capture of follow-up data and outcomes, it is possible

additional data lag is still present, challenging the completeness and accuracy of outcome assessment.

Due to HIPAA compliance protection, patients above 89 years old were included as a single category of age in the dataset; with age being an important risk predictor of mortality, this can potentially lead to some performance degradation for patients above 89 years old. Owing to the lack of symptom information and the use of oxygen therapies in the dataset, model performance can be potentially further improved if it is made available. Similarly, this negatively impacts the evaluation of existing prognostic scores which require FiO_2. We have referred to the best currently available information of clinical trial for threshold calculation, there could still exist differences in patient population between the RECOVERY trial and this current work. Additional work is required for validating the results on vaccinated population.

Preprint JMIR Publications

REFERENCES:

1. Knight SR, Ho A, Pius R, Buchan I, Carson G, Drake TM, Dunning J, Fairfield CJ, Gamble C, Green CA, Gupta R, Halpin S, Hardwick HE, Holden KA, Horby PW, Jackson C, McLean KA, Merson L, Nguyen-Van-Tam JS, Norman L, Noursadeghi M, Olliaro PL, Pritchard MG, Russell CD, Shaw CA, Sheikh A, Solomon T, Sudlow C, Swann O V., Turtle LCW, Openshaw PJM, Baillie JK, Semple MG, Docherty AB, Harrison EM. Risk stratification of patients admitted to hospital with covid-19 using the ISARIC WHO Clinical Characterisation Protocol: Development and validation of the 4C Mortality Score. BMJ 2020;370(September):1–13. PMID:32907855
2. Liang W, Liang H, Ou L, Chen B, Chen A, Li C, Li Y, Guan W, Sang L, Lu J, Xu Y, Chen G, Guo H, Guo J, Chen Z, Zhao Y, Li S, Zhang N, Zhong N, He J. 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(8):1081–1089. PMID:32396163
3. Knaus WA, Draper EA, Wagner DP, Zimmerman JE. APACHE II: a severity of disease classification system. Crit Care Med United States; 1985 Oct;13(10):818–829. PMID:3928249
4. Vincent JL, Moreno R, Takala J, Willatts S, De Mendonça A, Bruining H, Reinhart CK, Suter PM, Thijs LG. The SOFA (Sepsis-related Organ Failure Assessment) score to describe organ dysfunction/failure. On behalf of the Working Group on Sepsis-Related Problems of the European Society of Intensive Care Medicine. Intensive Care Med. United States; 1996. p. 707–710. PMID:8844239
5. Chen JH, Chang SS, Liu JJ, Chan RC, Wu JY, Wang WC, Lee SH, Lee CC. Comparison of clinical characteristics and performance of pneumonia severity score and CURB-65 among younger adults, elderly and very old subjects. Thorax 2010;65(11):971–977. [doi: 10.1136/thx.2009.129627]
6. Zou X, Li S, Fang M, Hu M, Bian Y, Ling J, Yu S, Jing L, Li D, Huang J. Acute Physiology

and Chronic Health Evaluation II Score as a Predictor of Hospital Mortality in Patients of Coronavirus Disease 2019. Crit Care Med 2020;48(8):E657–E665. PMID:32371611

7. Yadaw AS, Li Y chak, Bose S, Iyengar R, Bunyavanich S, Pandey G. Clinical features of COVID-19 mortality: development and validation of a clinical prediction model. Lancet Digit Heal [Internet] The Author(s). Published by Elsevier Ltd. This is an Open Access article under the CC BY-NC-ND 4.0 license; 2020;2(10):e516–e525. [doi: 10.1016/S2589-7500(20)30217-X]
8. Hajian-Tilaki K. Sample size estimation in diagnostic test studies of biomedical informatics. J Biomed Inform [Internet] Elsevier Inc.; 2014;48:193–204. PMID:24582925
9. van Buuren S, Groothuis-Oudshoorn K. mice: Multivariate Imputation by Chained Equations in R. J Stat Software, Artic [Internet] 2011;45(3):1–67. [doi: 10.18637/jss.v045.i03]
10. Chen T, Guestrin C. XGBoost: A scalable tree boosting system. Proc ACM SIGKDD Int Conf Knowl Discov Data Min 2016;13-17-Augu:785–794. [doi: 10.1145/2939672.2939785]
11. Li WT, Ma J, Shende N, Castaneda G, Chakladar J, Tsai JC, Apostol L, Honda CO, Xu J, Wong LM, Zhang T, Lee A, Gnanasekar A, Honda TK, Kuo SZ, Yu MA, Chang EY, Rajasekaran MR, Ongkeko WM. Using machine learning of clinical data to diagnose COVID-19: A systematic review and meta-analysis. BMC Med Inform Decis Mak BMC Medical Informatics and Decision Making; 2020;20(1):1–13. PMID:32993652
12. Yu C, Lei Q, Li W, Wang X, Liu W, Fan X, Li W. Clinical Characteristics, Associated Factors, and Predicting COVID-19 Mortality Risk: A Retrospective Study in Wuhan, China. Am J Prev Med [Internet] 2020;59(2):168–175. [doi: https://doi.org/10.1016/j.amepre.2020.05.002]
13. Patrício A, Costa RS, Henriques R. Predictability of COVID-19 Hospitalizations, Intensive Care Unit Admissions, and Respiratory Assistance in Portugal: Longitudinal Cohort Study. J Med Internet Res [Internet] 2021 Apr;23(4):e26075. [doi: 10.2196/26075]
14. Vaid A, Jaladanki SK, Xu J, Teng S, Kumar A, Lee S, Somani S, Paranjpe I, De Freitas JK,

Wanyan T, Johnson KW, Bicak M, Klang E, Kwon YJ, Costa A, Zhao S, Miotto R, Charney AW, Böttinger E, Fayad ZA, Nadkarni GN, Wang F, Glicksberg BS. Federated Learning of Electronic Health Records Improves Mortality Prediction in Patients Hospitalized with COVID-19. medRxiv Prepr Serv Heal Sci. 2020. PMID:32817979

15. Cheng F-Y, Joshi H, Tandon P, Freeman R, Reich DL, Mazumdar M, Kohli-Seth R, Levin MA, Timsina P, Kia A. Using Machine Learning to Predict ICU Transfer in Hospitalized COVID-19 Patients. J Clin Med [Internet] 2020;9(6). [doi: 10.3390/jcm9061668]
16. Parchure P, Joshi H, Dharmarajan K, Freeman R, Reich DL, Mazumdar M, Timsina P, Kia A. Development and validation of a machine learning-based prediction model for near-term in-hospital mortality among patients with COVID-19. BMJ Support \& Palliat Care [Internet] British Medical Journal Publishing Group; 2020; [doi: 10.1136/bmjspcare-2020-002602]
17. Li Z, Wang L, Huang L shuai, Zhang M, Cai X, Xu F, Wu F, Li H, Huang W, Zhou Q, Yao J, Liang Y, Liu G. Efficient management strategy of COVID-19 patients based on cluster analysis and clinical decision tree classification. Sci Rep [Internet] Nature Publishing Group UK; 2021;11(1):1–13. PMID:33953307
18. Ke G, Meng Q, Finley T, Wang T, Chen W, Ma W, Ye Q, Liu TY. LightGBM: A highly efficient gradient boosting decision tree. Adv Neural Inf Process Syst 2017;2017-Decem(Nips):3147–3155.
19. Guyon I, Weston J, Barnhill S, Vapnik V. Gene Selection for Cancer Classification using Support Vector Machines. Mach Learn [Internet] 2002;46(1):389–422. [doi: 10.1023/A:1012487302797]
20. Brier GW. Verification of forecasts expressed in terms of probability. Mon Weather Rev [Internet] Boston MA, USA: American Meteorological Society; 1950;78(1):1–3. [doi: 10.1175/1520-0493(1950)078<0001:VOFEIT>2.0.CO;2]
21. Niculescu-Mizil A, Caruana R. Predicting Good Probabilities with Supervised Learning. Proc

22nd Int Conf Mach Learn [Internet] New York, NY, USA: Association for Computing Machinery; 2005. p. 625–632. [doi: 10.1145/1102351.1102430]

22. Goodhue DL, Lewis W, Thompson R. Does PLS Have Advantages for Small Sample Size or Non-Normal Data? MIS Q [Internet] Management Information Systems Research Center, University of Minnesota; 2012 Jun 17;36(3):981–1001. [doi: 10.2307/41703490]

23. Stephens JR, Stümpfle R, Patel P, Brett S, Broomhead R, Baharlo B, Soni S. Analysis of Critical Care Severity of Illness Scoring Systems in Patients with Coronavirus Disease 2019: A Retrospective Analysis of Three U.K. ICUs. Crit Care Med 2020;E105–E107. PMID:32991357

24. Capelastegui A, España PP, Quintana JM, Areitio I, Gorordo I, Egurrola M, Bilbao A. Validation of a predictive rule for the management of community-acquired pneumonia. Eur Respir J 2006;27(1):151–157. PMID:16387948

25. Wellbelove Z, Walsh C, Perinpanathan T, Lillie P, Barlow G. Comparing the 4C mortality score for COVID-19 to established scores (CURB65, CRB65, qSOFA, NEWS) for respiratory infection patients. J Infect. 2021. p. 414–451. PMID:33115655

26. Liu JL, Xu F, Zhou H, Wu XJ, Shi LX, Lu RQ, Farcomeni A, Venditti M, Zhao YL, Luo SY, Dong XJ, Falcone M. Expanded CURB-65: A new score system predicts severity of community-acquired pneumonia with superior efficiency. Sci Rep [Internet] Nature Publishing Group; 2016;6(March):1–7. PMID:26987602

27. National Early Warning Score (NEWS) 2 | RCP London [Internet]. [cited 2021 Mar 29]. Available from: https://www.rcplondon.ac.uk/projects/outputs/national-early-warning-score-news-2

28. Carr E, Bendayan R, Bean D, O'Gallagher K, Pickles A, Stahl D, Zakeri R, Searle T, Shek A, Kraljevic Z, Teo JT, Shah A, Dobson R, Gallagher K, Bean D, Pickles A, Stahl D, Zakeri R, Searle T, Shek A, Kraljevic Z, Teo JT, Shah A, Dobson R. Evaluation and Improvement of the

National Early Warning Score (NEWS2) for COVID-19: a multi-hospital study Adding. medRxiv BMC Medicine; 2020;2020.04.24.20078006.

29. Prower E, Grant D, Bisquera A, Breen CP, Camporota L, Gavrilovski M, Pontin M, Douiri A, Glover GW. The ROX index has greater predictive validity than NEWS2 for deterioration in Covid-19. EClinicalMedicine 2021;35. [doi: 10.1016/j.eclinm.2021.100828]

30. Roca O, Messika J, Caralt B, García-de-Acilu M, Sztrymf B, Ricard J-D, Masclans JR. Predicting success of high-flow nasal cannula in pneumonia patients with hypoxemic respiratory failure: The utility of the ROX index. J Crit Care [Internet] 2016;35:200–205. [doi: https://doi.org/10.1016/j.jcrc.2016.05.022]

31. Lundberg SM, Erion G, Chen H, DeGrave A, Prutkin JM, Nair B, Katz R, Himmelfarb J, Bansal N, Lee S-I. From local explanations to global understanding with explainable AI for trees. Nat Mach Intell [Internet] 2020;2(1):56–67. [doi: 10.1038/s42256-019-0138-9]

32. Lundberg SM, Lee S-I. A Unified Approach to Interpreting Model Predictions. In: Guyon I, Luxburg U V, Bengio S, Wallach H, Fergus R, Vishwanathan S, Garnett R, editors. Adv Neural Inf Process Syst [Internet] Curran Associates, Inc.; 2017. Available from: https://proceedings.neurips.cc/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf

33. Habibzadeh F, Habibzadeh P, Yadollahie M. On determining the most appropriate test cut-off value: The case of tests with continuous results. Biochem Medica 2016;26(3):297–307. PMID:27812299

34. YOUDEN WJ. Index for rating diagnostic tests. Cancer United States; 1950 Jan;3(1):32–35. PMID:15405679

35. Schisterman EF, Perkins NJ, Liu A, Bondell H. Optimal cut-point and its corresponding Youden Index to discriminate individuals using pooled blood samples. Epidemiology United States; 2005 Jan;16(1):73–81. PMID:15613948

36. Shapiro DE. The interpretation of diagnostic tests. Stat Methods Med Res England; 1999 Jun;8(2):113–134. PMID:10501649

37. Greiner M, Pfeiffer D, Smith R.D. Principles and practical application of the receiver-operating characteristic analysis for diagnostic tests. Prev Vet Med 2000;45:23–41. [doi: https://doi.org/10.1016/S0167-5877(00)00115-XGet]

38. Vickers AJ, Van Calster B, Steyerberg EW. Net benefit approaches to the evaluation of prediction models, molecular markers, and diagnostic tests. BMJ 2016;352. PMID:26810254

39. Vickers AJ, van Calster B, Steyerberg EW. A simple, step-by-step guide to interpreting decision curve analysis. Diagnostic Progn Res Diagnostic and Prognostic Research; 2019;3(1):1–8. PMID:31592444

40. Subudhi S, Verma A, Patel AB, Hardin CC, Khandekar MJ, Lee H, McEvoy D, Stylianopoulos T, Munn LL, Dutta S, Jain RK. Comparing machine learning algorithms for predicting ICU admission and mortality in COVID-19. npj Digit Med [Internet] Springer US; 2021;4(1):1–7. [doi: 10.1038/s41746-021-00456-x]

41. Bolourani S, Brenner M, Wang P, McGinn T, Hirsch JS, Barnaby D, Zanos TP. A Machine Learning Prediction Model of Respiratory Failure Within 48 Hours of Patient Admission for COVID-19: Model Development and Validation. J Med Internet Res [Internet] 2021 Feb;23(2):e24246. [doi: 10.2196/24246]

42. Kim H-J, Han D, Kim J-H, Kim D, Ha B, Seog W, Lee Y-K, Lim D, Hong SO, Park M-J, Heo J. An Easy-to-Use Machine Learning Model to Predict the Prognosis of Patients With COVID-19: Retrospective Cohort Study. J Med Internet Res [Internet] 2020 Nov;22(11):e24225. [doi: 10.2196/24225]

43. Challen R, Brooks-Pollock E, Read JM, Dyson L, Tsaneva-Atanasova K, Danon L. Risk of mortality in patients infected with SARS-CoV-2 variant of concern 202012/1: Matched cohort study. BMJ 2021;372:1–10. PMID:33687922

44. O'Driscoll M, Ribeiro Dos Santos G, Wang L, Cummings DAT, Azman AS, Paireau J, Fontanet A, Cauchemez S, Salje H. Age-specific mortality and immunity patterns of SARS-CoV-2. Nature [Internet] 2021;590(7844):140–145. [doi: 10.1038/s41586-020-2918-0]

45. Zhou J, Lee S, Wang X, Li Y, Wu WKK, Liu T, Cao Z, Zeng DD, Leung KSK, Wai AKC, Wong ICK, Cheung BMY, Zhang Q, Tse G. Development of a multivariable prediction model for severe COVID-19 disease: a population-based study from Hong Kong. npj Digit Med [Internet] 2021;4(1):66. [doi: 10.1038/s41746-021-00433-4]

46. Ok F, Erdogan O, Durmus E, Carkci S, Canik A. Predictive values of blood urea nitrogen/creatinine ratio and other routine blood parameters on disease severity and survival of COVID-19 patients. J Med Virol 2021;93(2):786–793. PMID:32662893

47. Cheng A, Hu L, Wang Y, Huang L, Zhao L, Zhang C, Liu X, Xu R, Liu F, Li J, Ye D, Wang T, Lv Y, Liu Q. Diagnostic performance of initial blood urea nitrogen combined with D-dimer levels for predicting in-hospital mortality in COVID-19 patients. Int J Antimicrob Agents Elsevier Ltd; 2020;56(3). PMID:32712332

48. Marjot T, Webb GJ, Barritt AS, Moon AM, Stamataki Z, Wong VW, Barnes E. COVID-19 and liver disease: mechanistic and clinical perspectives. Nat Rev Gastroenterol Hepatol [Internet] Springer US; 2021;18(5):348–364. PMID:33692570

49. Razavian N, Major VJ, Sudarshan M, Burk-Rafel J, Stella P, Randhawa H, Bilaloglu S, Chen J, Nguy V, Wang W, Zhang H, Reinstein I, Kudlowitz D, Zenger C, Cao M, Zhang R, Dogra S, Harish KB, Bosworth B, Francois F, Horwitz LI, Ranganath R, Austrian J, Aphinyanaphongs Y. A validated, real-time prediction model for favorable outcomes in hospitalized COVID-19 patients. npj Digit Med [Internet] Springer US; 2020;3(1). [doi: 10.1038/s41746-020-00343-x]

50. Liu H, Chen J, Yang Q, Lei F, Zhang C, Qin J-J, Chen Z, Zhu L, Song X, Bai L, Huang X, Liu W, Zhou F, Chen M-M, Zhao Y-C, Zhang X-J, She Z-G, Xu Q, Ma X, Zhang P, Ji Y-X, Zhang X, Yang J, Xie J, Ye P, Azzolini E, Aghemo A, Ciccarelli M, Condorelli G, Stefanini GG, Xia

J, Zhang B-H, Yuan Y, Wei X, Wang Y, Cai J, Li H. Development and validation of a risk score using complete blood count to predict in-hospital mortality in COVID-19 patients. Med [Internet] Elsevier Inc.; 2021;2(4):435-447.e4. [doi: 10.1016/j.medj.2020.12.013]

51. Bi X, SU Z, Yan H, Du J, Wang J, Chen L, Peng M, Chen S, Shen B, Li J. Prediction of severe illness due to COVID-19 based on an analysis of initial Fibrinogen to Albumin Ratio and Platelet count. Platelets [Internet] Taylor & Francis; 2020;31(5):674–679. PMID:32367765

52. Kheir M, Saleem F, Wang C, Mann A, Chua J. Higher albumin levels on admission predict better prognosis in patients with confirmed COVID-19. PLoS One [Internet] Public Library of Science; 2021 Mar 16;16(3):e0248358. Available from: https://doi.org/10.1371/journal.pone.0248358

53. Natarajan A, Su HW, Heneghan C. Assessment of physiological signs associated with COVID-19 measured using wearable devices. npj Digit Med [Internet] Springer US; 2020;3(1). [doi: 10.1038/s41746-020-00363-7]

54. The RECOVERY Collaborative Group. Dexamethasone in Hospitalized Patients with Covid-19. N Engl J Med 2021;384(8):693–704. PMID:32678530

55. W L, M VDE, R L, W B, N K, G T, S L, J M. Defining community acquired pneumonia severity on presentation to hospital: An international derivation and validation study. Thorax 2003;58(5):377–382.

56. Kaysin A, Viera AJ. Community-Acquired Pneumonia in Adults: Diagnosis and Management. Am Fam Physician United States; 2016 Nov;94(9):698–706. PMID:27929242

ACKNOWLEDGEMENTS

We would like to acknowledge the extensive programming and planning work of the Amgen Center for Observational Research (Oana Abrahamian, Bagmeet Behera, Corinne Brooks and Kimberly A. Roehl) and initial modeling and feasibility analysis by Amgen's Digital Health & Innovation (Data Science and Engineering team). We also would like to acknowledge the health care professionals whose tireless efforts in this unprecedented pandemic have provided critical knowledge, as well as the patients from whom we continue to learn so much.

AUTHOR CONTRIBUTIONS

All the authors participated in literature search, conceptualization, data interpretation, reviewing and editing the manuscript. JP and FH were responsible for study design and methodology; FH performed the formal analysis and produced formatted tables and figures, and the original draft. AM conducted initial feasibility analysis. All the authors have full access to the data in the study and accept responsibility to submit for publication.

COMPETING INTERESTS

Fang He, John H. Page, Anirban Mishra are employees and stockholders of Amgen, Inc. Kerry Weinberg, an employee of League Inc., was formerly an employee of Amgen, Inc. and owns stock in Amgen, Inc.

Preprint JMIR Publications

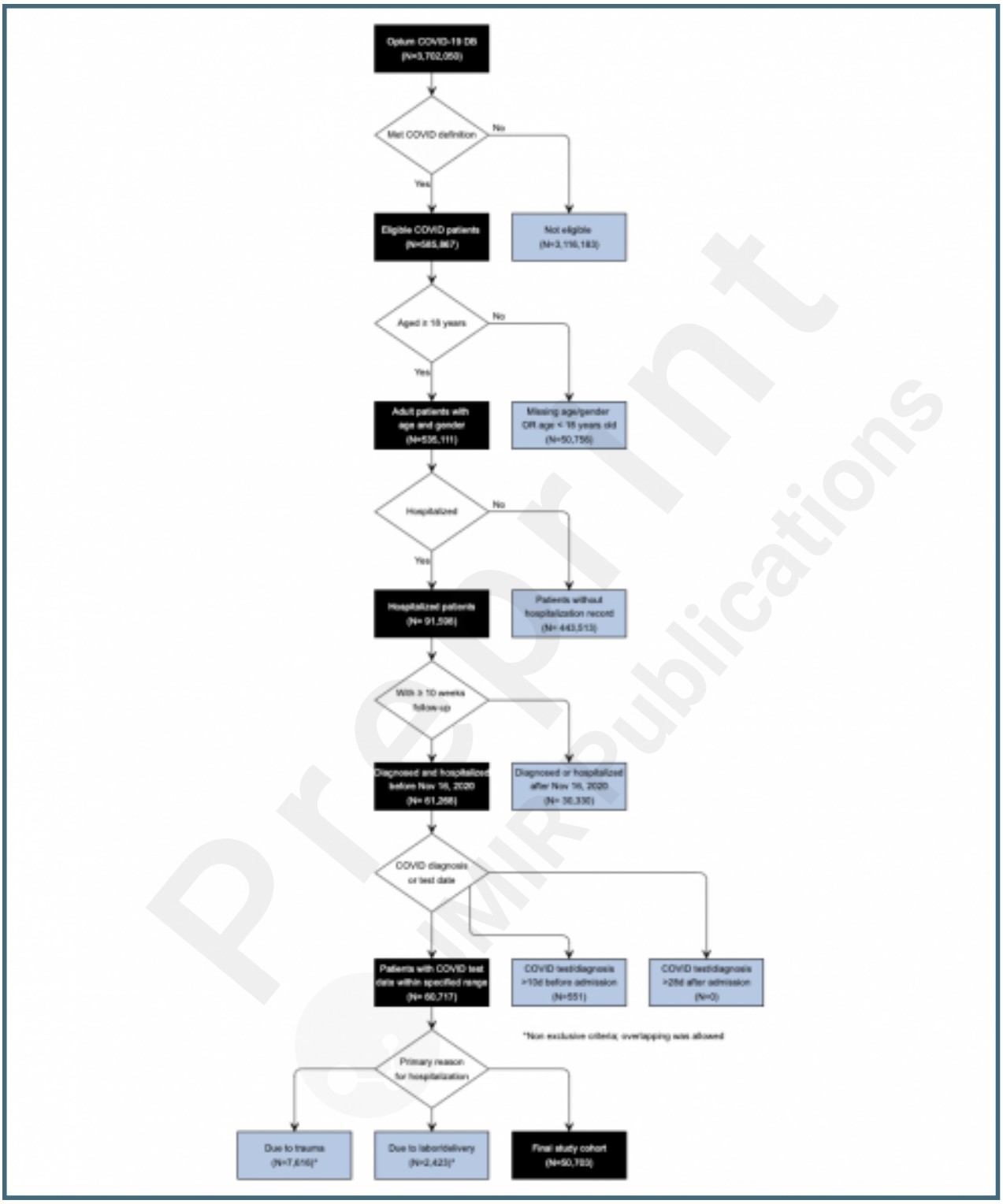
Supplementary Files

Preprint JMIR Publications

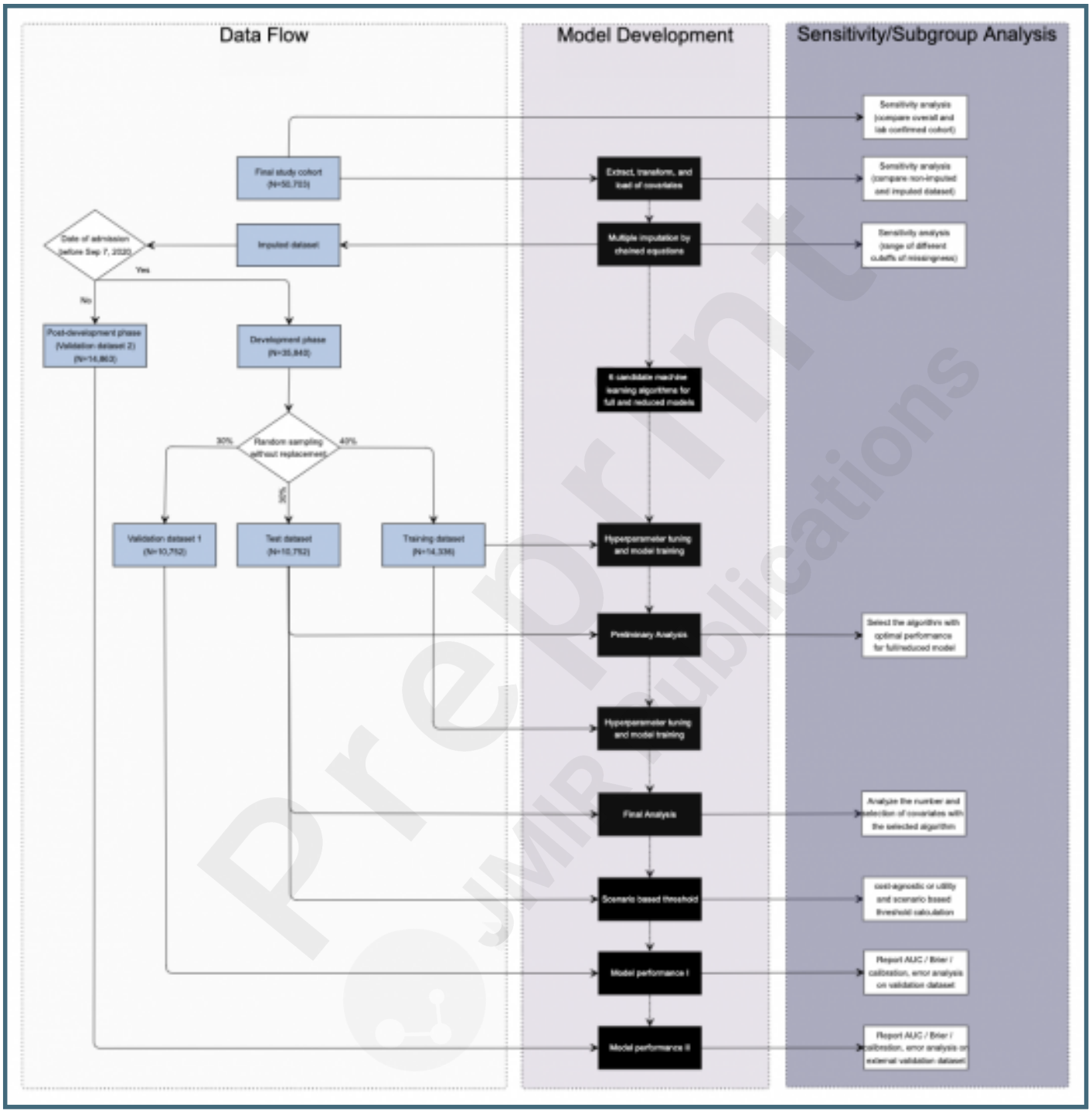
Figures

Preprint JMIR Publications

Patient attrition diagram.

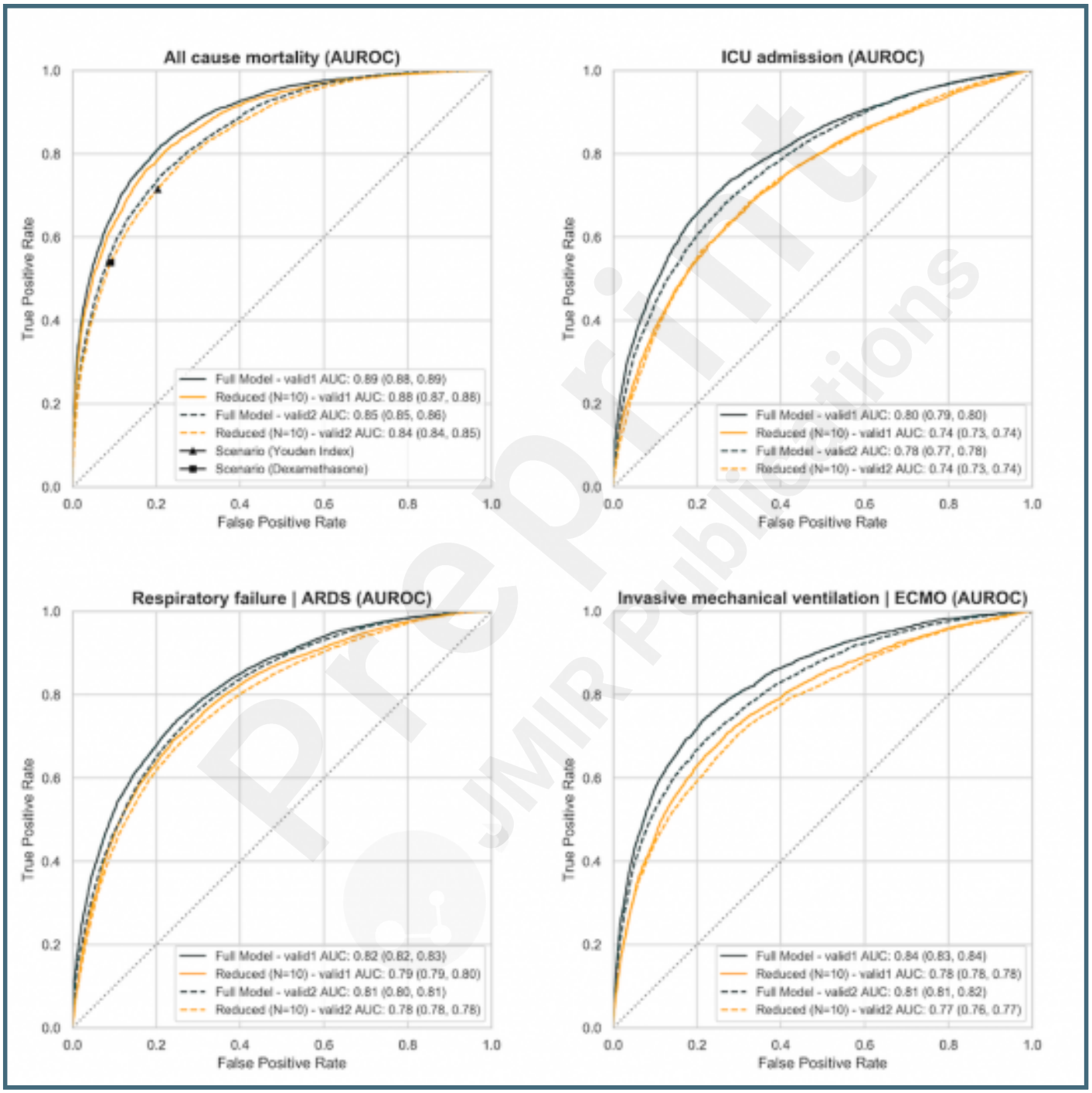


Model development and validation framework including data sampling and corresponding sensitivity analyses.

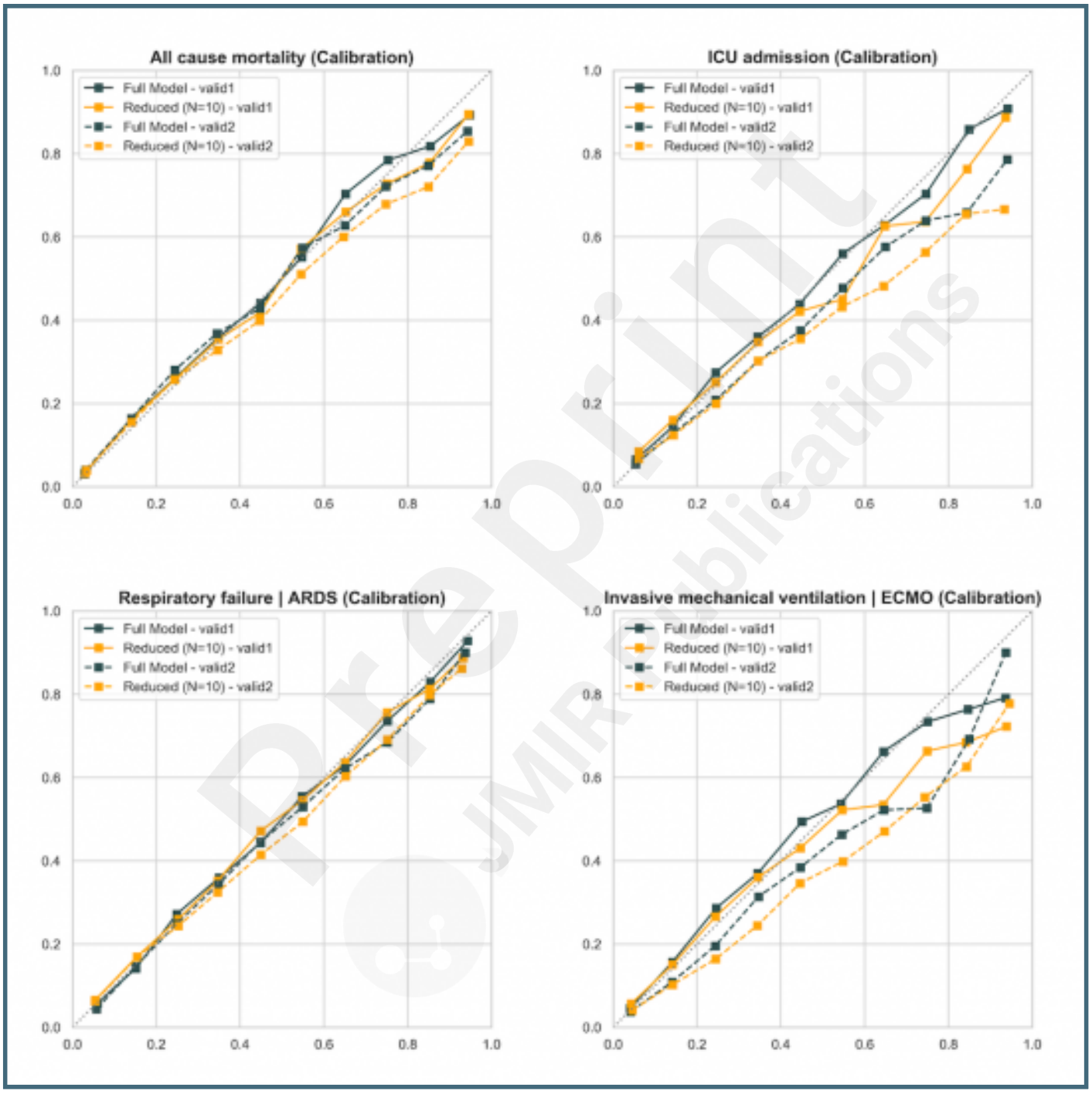


Preprint JMIR Publications

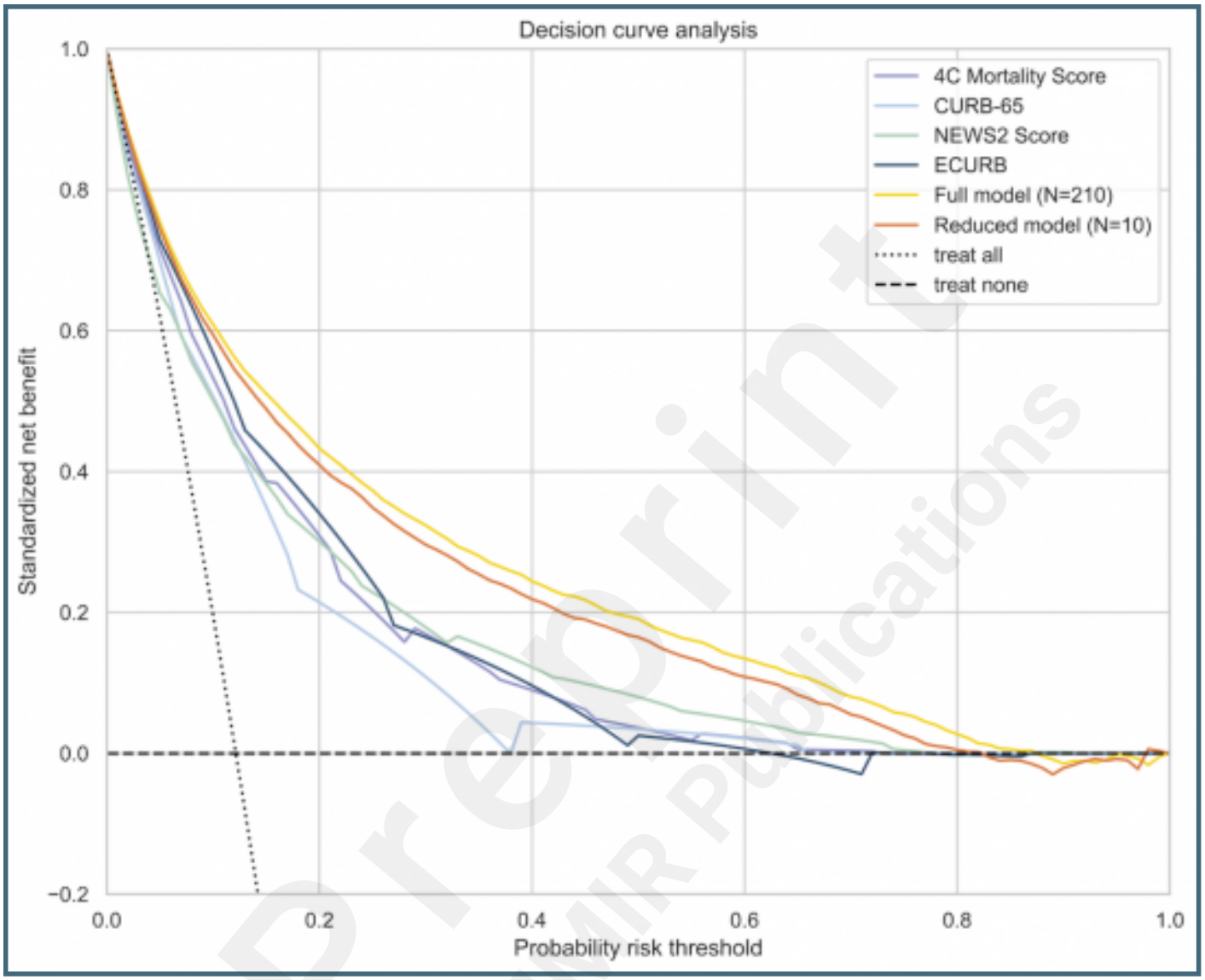
Receiver operating characteristics (AUROC) curves on four prediction outcomes: (a) all-cause mortality; (b) respiratory failure including ARDS; (c) ICU admission; (d) invasive mechanical ventilation including ECMO. Full model is colored in black, parsimonious model with ten input variables is colored in orange. Solid line represents result from validation 1 during model development phase (N=10,752); dashed line represents result from independent validation 2 from post-development phase (N=14,336)).



Calibration curve (number of bins = 10) on four prediction outcomes: (a) all-cause mortality; (b) respiratory failure including ARDS; (c) ICU admission; (d) invasive mechanical ventilation including ECMO. Full model is colored in black, parsimonious model with ten input variables is colored in orange. Solid line represents result from validation 1 during model development phase (N=10,752); dashed line represents result from independent validation 2 from post-development phase (N=14,336)).



Decision curve analysis of standardized net benefit across different risk thresholds. Dotted line represents the scenario if everyone is treated; dashed line represents the scenario if none is treated.



Preprint JMIR Publications

CONSORT (or other) checklists

Preprint JMIR Publications

STARD checklist.
URL: http://asset.jmir.pub/assets/288593e73d4184f6d3b5d47a9892a4fa.pdf

Preprint
JMIR Publications