Development and Validation of a Machine learning prediction model of respiratory failure within 48 hours of patient admission for COVID-19

Siavash Bolourani, Max Brenner, Ping Wang, Thomas McGinn, Jamie Hirsch, Douglas Barnaby, Theodoros Zanos, Northwell COVID-19 Research Consortium

Submitted to: Journal of Medical Internet Research
on: September 10, 2020

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

Development and Validation of a Machine learning prediction model of respiratory failure within 48 hours of patient admission for COVID-19

Siavash Bolourani[1] MD; Max Brenner[1] MD, PhD; Ping Wang[1] MD; Thomas McGinn[1] MD, MPH; Jamie Hirsch[1] MD; Douglas Barnaby[1*] MD; Theodoros Zanos[1*] PhD, MSc, BEng; Northwell COVID-19 Research Consortium[1]

[1]Feinstein Institutes for Medical Research Northwell Health Manhasset US
[*]these authors contributed equally

Corresponding Author:
Theodoros Zanos PhD, MSc, BEng
Feinstein Institutes for Medical Research
Northwell Health
350 Community Dr
Room 1257
Manhasset
US

Abstract

Background: Predicting respiratory failure in COVID-19 patients based on an early clinical profile can help triaging, resource allocation, and morbidity reduction by appropriately monitoring patients at risk. Given the complexity of the disease, this effort can benefit from machine learning (ML) approaches.

Objective: Our objective is to establish a machine learning model that predicts respiratory failure within 48 hours of admission based on data from the emergency department (ED).

Methods: Data was collected from patients with COVID-19 who were admitted to Northwell Health acute care hospitals and discharged, died, or spent a minimum of 48 hours in the hospital between March 1, 2020 and May 11, 2020. Of 11,525 patients, 933 (8.1%) were placed on invasive mechanical ventilation within 48 hours of admission. The variables used by the models included clinical and laboratory data commonly collected in the ED. We trained and validated an XgBoost model alongside two other prediction models using cross hospitals validation. We compared model performance and a Modified Early Warning Score (MEWS) using receiver operating characteristic (ROC) curves, precision-recall (PR) curves, and other metrics.

Results: The XgBoost model had the highest mean accuracy of 0.908 (AUC = 0.83), outperforming MEWS and the other models. Important variables included the type of oxygen delivery used in the ED, patient age, emergency severity score, respiratory rate, serum lactate, heart rate, and serum glucose values.

Conclusions: XgBoost has high predictive accuracy, outperforming other early warning scores. The clinical plausibility and predictive ability of XgBoost suggest that the model could be used to predict 48-hour respiratory failure in admitted patients with COVID-19.

(JMIR Preprints 10/09/2020:24246)
DOI: https://doi.org/10.2196/preprints.24246

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

Original Manuscript

Development and Validation of a Machine learning prediction model of respiratory failure within 48 hours of patient admission for COVID-19

Short running head: Predicting respiratory failure in COVID-19

Siavash Bolourani[1,2,3,4] MD, Max Brenner [2,4], MD PhD, Ping Wang [2,3,4,5] MD, Thomas McGinn[5,6] MD MPH, Jamie S. Hirsch[5,6,7] MD, Douglas P. Barnaby[*5,6] MD, Theodoros P. Zanos[*1,3,5] PhD
And the Northwell COVID-19 Research Consortium: Matthew Barish[5,6] MD, Stuart L. Cohen[5,6], MD; Kevin Coppa[7], BS, Karina W. Davidson [5,6] PhD, Shubham Debnath[1] PhD, Lawrence Lau[3,5] MD, Todd J. Levy[1] MS, Alexander Makhnevich[5] MD, Marc D. Paradis [8] SM, Viktor Tóth[1] MSc
**Co-senior authors
[1]*Institute of Bioelectronic Medicine, Feinstein Institutes for Medical Research, Northwell Health, Manhasset, NY 11030*
[2]*Center for Immunology and Inflammation, Feinstein Institutes for Medical Research, Northwell Health, Manhasset, NY 11030*
[3]*Elmezzi Graduate School of Molecular Medicine, Manhasset, NY 11030*
[4]*Department of Surgery, Donald and Barbara Zucker School of Medicine at Hofstra/Northwell, Manhasset, NY 11030.*
[5]*Donald and Barbara Zucker School of Medicine at Hofstra/Northwell, Northwell Health, Hempstead, NY*
[6]*Institute of Health Innovations and Outcomes Research, Feinstein Institutes for Medical Research, Northwell Health, Manhasset, NY*
[7]*Department of Information Services, Northwell Health, New Hyde Park, NY*
[8]*Department of Data Strategy & Ventures, Northwell Health, Manhasset, NY*

To whom correspondence should be addressed:
Theodoros P. Zanos, PhD
Assistant Professor
The Feinstein Institutes of Medical Research
Northwell Health
350 Community Drive, Room 1257
Manhasset, NY 11030
T: 516-562-0484
E: tzanos@northwell.edu

Abstract

Background: Predicting early respiratory failure in COVID-19 can help triage patients to higher levels of care, allocate scarce resources, and reduce morbidity and mortality by appropriately monitoring and treating patients at greatest risk for deterioration. Given the complexity of COVID-19 disease, machine learning (ML) approaches may support clinical decision making for patients with this disease.

Objective: Our objective is to derive a machine learning model that predicts respiratory failure within 48 hours of admission based on data from the emergency department (ED).

Methods: Data was collected from patients with COVID-19 who were admitted to Northwell Health acute care hospitals and discharged, died, or spent a minimum of 48 hours in the hospital between March 1, 2020 and May 11, 2020. Of 11,525 patients, 933 (8.1%) were placed on invasive mechanical ventilation within 48 hours of admission. Variables used by the models included clinical and laboratory data commonly collected in the ED. We trained and validated three predictive models (two based on XGBoost, one that utilized logistic regression) using cross hospitals validation. We compared model performance between all three models as well as an established early warning score (Modified Early Warning Score (MEWS)) using receiver operating characteristic (ROC) curves, precision-recall (PR) curves, and other metrics.

Results: The XGBoost model had the highest mean accuracy of 0.919 (AUC = 0.77), outperforming the other two models as well as MEWS. Important predictor variables included the type of oxygen delivery used in the ED, patient age, Emergency Severity Index (ESI), respiratory rate, serum lactate, and demographic characteristics.

Conclusions: XGBoost has high predictive accuracy, outperforming other early warning scores. The clinical plausibility and predictive ability of XGBoost suggest that the model could be used to predict 48-hour respiratory failure in admitted patients with COVID-19.

Key words: artificial intelligence; prognostic models; pandemic; severe acute respiratory syndrome coronavirus 2

Introduction

On March 11, 2020, coronavirus disease 2019 (COVID-19) due to SARS-CoV-2 infection was declared a pandemic by the World Health Organization [1]. As of December 16, 2020, more than 17 million people have been documented as infected and over 300,000 have died in the United States. During the first wave, New York lied at the epicenter of the pandemic in the United States, with over 390,000 cases and 30,000 deaths before the summer [2].

Respiratory failure is the leading cause of death among patients with COVID-19, with up to one-third of patients admitted with COVID-19 requiring invasive mechanical ventilation (IMV) [3–8]. The decision to initiate IMV in these patients is not straightforward. Many patients with severe disease appear comfortable despite profound hypoxemia, and they are commonly managed with supplemental oxygen, self-proning, and close monitoring [9,10]. However, some of these patients subsequently deteriorate and require IMV following transfer from the emergency department (ED). This subgroup has worse outcomes than those placed on IMV initially [11]. Before the surge of COVID-19, patients initially admitted to a non-critical care setting who needed an unplanned transfer to an intensive care unit (ICU) had greater morbidity and mortality than those admitted directly to a critical care unit [12–14]. Thus, accurately identifying patients at high risk for deterioration could improve clinical outcomes with closer monitoring, direct admission to a critical care unit, or earlier discussions regarding patient preferences and goals of care.

Identifying patients at high risk for, or in the early stages of, clinical deterioration has been actively researched for decades. The field has generated many severity-of-illness calculators, early warning scores, and, more recently, predictive analytic tools that use advanced machine learning and artificial intelligence [15–23]. Our goal was to derive a prediction model that estimates the risk of short-term (< 48 hours) respiratory failure for patients with COVID-19 who were not initially placed on IMV. Such a tool could improve outcomes by avoiding delayed admission to a critical care unit, providing additional respiratory support and closer monitoring, or initiating earlier discussions around the goals of care.

Methods

This retrospective observational cohort drew data from 13 acute care hospitals of Northwell Health, the largest health care system in New York State. Data was extracted from the electronic health record (EHR) Sunrise Clinical Manager (Allscripts, Chicago, IL). EHRs were screened for adult (≥ 21 years old) patients who were given a positive result of SARS-CoV-2 based on a nasopharyngeal sample tested by polymerase chain reaction. Included patients were hospitalized and discharged, died, or spent a minimum of 48 hours in the hospital between March 1, 2020 and May 11, 2020. For patients who had multiple qualifying hospital admissions, only the first hospitalization was included. Patients who were transferred between hospitals within the health system were treated as one hospital encounter. Total of 11,919 patients were identified. Patients were excluded if they were placed on mechanical ventilation prior to inpatient admission. A total of 11,525 patients remained for analysis. The Institutional Review Board of Northwell Health approved the study protocol and waived the requirement for informed consent.

Data Acquisition

Data collected from the EHR included patient demographics, comorbidities, home medications, initial vitals and laboratory values, treatments (e.g., oxygen therapy, mechanical ventilation), and clinical outcomes (e.g., length of stay, discharge, mortality). Vitals and laboratory testing were restricted to those obtained while the patient was in the ED.

Outcomes

The target outcome variable was defined as intubation and mechanical ventilation within 48 hours of admission. In the EHR, the admission time was recorded, and the intubation event was defined as the first time mechanical ventilation was recorded.

Predictive Machine Learning Model

We evaluated three predictive models: XGBoost, XGBoost + SMOTEENN (combined oversampling using SMOTE and undersampling using Edited Nearest Neighbours) [24], and Logistic Regression [25]. XGBoost combines a recursive gradient–boosting method, called Newton boosting, with a decision-tree model. Given that each tree is boosted in parallel, the model efficiently provides accurate predictions [26] . Furthermore, because each tree is boosted recursively and in parallel, the model benefits from high interpretability of the variable importance features.

XGBoost + SMOTEENN method involves combined over- and under-sampling using SMOTE and Edited Nearest Neighbors respectively on the training set before training a XGBoost model [27]. This method has been shown best performance in the resampling datasets [28]. Furthermore, in our experience, when using any of the oversampling or undersampling methods alone calibration of the model is severely affected. However, when we combine oversampling the minority class with undersampling of majority class, we found that we get a more accurate model both in discriminability and minimizing the effect on the calibration of the model.

For every learning framework, we validated the model with each hospital external validation (i.e. for each fold, one hospital was picked as a testing test and the others as a training set). Only hospitals with >1,000 covid-19 patients in the data were picked for the testing sets, and a random sample of 1,000 patients were picked to be our testing set for each fold. Grid search was used to hypertune the parameters of the respective models. The XGBoost model was tuned based on min_child_weight, gamma, subsample, colsample_bytree, and max_depth parameters, and the ranges of the values were 1-20, 0.5-20, 0.2-1.0, 0.2-1.0, and 2-40 respectively.

When data were missing, we imputed weighted k-nearest neighbors [29] for numerical values and added a category 'missing' for categorical values. We used one-hot [30] to encode categorical variables as a one-hot numeric array. The most important variables were calculated based on a decrease in the mean gini coefficient (i.e., the variables most useful in splitting the data to help make a prediction) for XGBoost and XGBoost + SMOTEENN; and by absolute value of regression coefficient for Logistic Regression, and were calculated based on the largest hospital testing set. The resulting receiver operating characteristic (ROC) curves and corresponding accuracy, recall (sensitivity), specificity, geometric mean, and F_β-Score were averaged. For the F_β-Score, the β parameter value was designated as $\beta = 4$ to capture a higher detriment of false negatives than false positives (i.e., if we value recall, β times as much the precision). For definitions of these measures and how they were calculated, see Multi.

Calibration curves (reliability curves) were plotted by dividing the testing sets (for each hospital fold) into 10 bins randomly with an increasing fraction of patients that had respiratory failure in the sample. The fraction positives (patients who had respiratory failure) and their mean corresponding predicted value from the corresponding model were depicted and averaged into 10 bins. The Brier score was calculated for each external hospital fold and the mean Brier score and standard deviation was calculated and depicted in the legend of the calibration curve. For further explanation of these measures and how they were calculated, see Multimedia Appendix 1.

Python 2.6 was used to implement our machine learning framework. Respective prediction models of XGBoost and Logistic Regression were used from the scikit-learn Application Programming

Interface (API) in Python [31]. GridSearchCV from the scikit-learn API was used to perform the grid search and hypertune the parameters. We used the default imblearn API version of the SMOTEENN [27]. SimpleImputer [32] was used for imputations, which were replaced with a new category missing. KNNImputer [33] was used to impute the missing numerical data [29]. The default value for k = 5 was not changed. OneHotEncoder from sklearn API was used to transform categorical variables to one-hot numeric arrays.

Modified Early Warning Score

The Modified Early Warning Score (MEWS) was computed from patient vital signs (Multimedia Appendix 2) and is a variant of other known and used risk scores [34,35]. MEWS ranges from 0 to 15 and incorporates heart rate (beats per minute), respiratory rate (breaths per minute), systolic blood pressure (mmHg), and body temperature (degrees Celsius). In our dataset, one MEWS subcomponent, the AVPU (alert, verbal, pain, unresponsive) neurologic assessment, had a significant amount of missing data (> 80%; data not shown) and was not included in the MEWS calculation for this project. An elevated MEWS score indicates a risk for clinical instability, including death or the need for ICU admission [36]. In 2012, our health system created a custom modification that was incorporated into the EHR. It includes automatic calculation and display via Arden Syntax Medical Logic Modules [37]. Based on local health-system guidelines, any score ≥ 7 requires an escalation in intensity of care. For example, MEWS > 7 requires increased frequency of vital sign measurement (every 2 hours), MEWS > 8 requires evaluation by a licensed independent provider, MEWS > 9 requires consideration of evaluation by a rapid response team, and MEWS > 10 requires a change in the level of service per a defined protocol. For the MEWS score, we chose the highest value while in the ED.

Results

Patient Characteristics

During the study period, we identified 11,525 patients admitted from the ED with a diagnosis of COVID-19. Of these, 933 (8.0%) were placed on IMV within 48 hours of admission. Baseline characteristics (demographics, baseline vital signs, and laboratory measurements) for all patients are shown in Table 1, stratified by study outcome. Comorbidities were captured from ICD-10 codes listed in the EHR.

Table 1. Demographic, clinical and laboratory data from hospitalized patients

	Not Intubated **n = 10,592**	**Intubated** **n = 933**	**Missing** **(%)**
Demographic Characteristics			
Age, y, median [IQR]	65.00 [54.00, 77.00]	66.00 [56.00, 75.00]	0
Female, n (%)	4,530 (42.8)	327 (35.0)	0
Primary language, English, n (%)	8,498 (80.2)	746 (80.0)	0
Race, n (%)			0
Black	2,199 (20.8)	236 (25.3)	
Asian	889 (8.4)	77 (8.3)	
White	4,148 (39.2)	310 (33.2)	
Declined	71 (0.7)	8 (0.9)	

Other	2,884 (27.2)	268 (28.7)	
Unknown	401 (3.8)	34 (3.6)	
Ethnicity, n (%)			0.1
Hispanic or Latino	2,238 (21.1)	202 (21.7)	
Not Hispanic or Latino	7685 (72.6)	648 (69.5)	
Declined	43 (0.4)	1 (0.1)	
Unknown	618 (5.8)	82 (8.8)	
Vital Signs			
SBP, mmHg, median [IQR]	134.00 [118.00, 150.00]	134.00 [115.00, 151.75]	0.5
DBP, mmHg, median [IQR]	79.00 [70.50, 87.00]	77.00 [69.00, 86.00]	0.6
HR, beats/min, median [IQR]	94.00 [85.00, 102.00]	97.00 [88.50, 112.00]	0.4
RR, breaths/min, median [IQR]	21.00 [18.00, 25.00]	24.00 [20.00, 32.00]	0.8
Temperature, °C, mean (SD)	37.77 (0.97)	37.86 (1.11)	1.6
Oxygen saturation, %, median [IQR]	97.00 [95.00, 98.00]	96.00 [93.00, 98.00]	1.7
BMI, mean (SD)	29.12 (7.79)	30.39 (9.21)	47.1
Laboratory Data			
White blood cell count, x10^9/L, median [IQR]	7.34 [5.45, 9.92]	8.25 [6.20, 11.50]	9
Absolute neutrophil count, x10^9/L, median [IQR]	5.68 [3.95, 8.11]	6.84 [4.76, 9.62]	11.5
Absolute Lymphocyte Count, x10^9/L, median [IQR]	0.90 [0.63, 1.27]	0.80 [0.56, 1.13]	11.5
Hemoglobin, g/dL, mean (SD)	12.93 (2.12)	13.14 (2.11)	9
Platelets, K/uL, mean (SD)	230.17 (101.93)	217.19 (87.45)	9.1
Sodium, mmol/L, mean (SD),	136.64 (6.21)	135.38 (5.74)	11.9
Carbon dioxide, mmol/L, mean (SD)	23.61 (3.79)	22.67 (4.68)	11.9
Creatinine, mg/dL, median [IQR]	1.03 [0.80, 1.46]	1.20 [0.92, 1.75]	12
Bilirubin, mg/dL, median [IQR]	0.50 [0.40, 0.70]	0.60 [0.40, 0.80]	12.5

Ferritin, ng/mL, mean (SD)	1283.50 (2732.65)	1731.05 (2631.38)	73.2
Procalcitonin, mean (SD), ng/mL	1.22 (10.96)	2.12 (8.16)	66.3
D-dimer, ng/mL, mean (SD)	1871.84 (5306.42)	2659.09 (6798.96)	65.4
Lactate dehydrogenase, U/L, mean (SD)	455.61 (213.04)	611.05 (272.16)	71
pH, arterial, mean (SD)	7.42 (0.09)	7.39 (0.11)	96.7
pO_2, arterial, mmHg, mean (SD)	99.90 (65.17)	85.26 (61.42)	94.8
pCO_2, arterial, mmHg, mean (SD)	34.66 (9.38)	35.38 (11.45)	94.7
Comorbidities			
Hypertension, n (%)	1183 (11.2)	115 (12.3)	0
Diabetes, n (%)	685 (6.5)	77 (8.3)	0
Coronary artery disease, n (%)	148 (1.4)	15 (1.6)	0
Asthma/COPD, n (%)	242 (2.3)	20 (2.1)	0
Chronic kidney disease, n (%)	99 (0.9)	8 (0.9)	0
HIV, n (%)	26 (0.2)	1 (0.1)	0

Definition of abbreviations: BMI = body mass index; COPD = chronic obstructive pulmonary disease; DBP = diastolic blood pressure; HR = heart rate; IQR = interquartile range; pCO_2 = partial pressure of carbon dioxide; pO_2= partial pressure of oxygen; RR = respiratory rate; SBP = systolic blood pressure; SD = standard deviation.

Prediction Models for Respiratory Failure

Based on XGBoost, the mean area under the curve (AUC) of the ROC curve was 0.77 ± 0.05 and the mean AUC of the PR curve (AUCPR) was 0.26 ± 0.04 (Figure 1). The 10 most important variables, in order of decreasing importance, were: most invasive mode of oxygen delivery being a non-rebreather mask, ESI values of 1 and 3, maximum respiratory rate, maximum oxygen saturation, black race, age on admission, eosinophil percentage, serum sodium level, and serum lactate level (Figure 1). The confusion matrix for the model's largest hospital testing set showed that most false predictions were false negatives (those who were predicted to not require intubation but were intubated within 48 hours). False positives (those who were predicted to require intubation but were not intubated within 48 hours) were the minority of predictions (Figure 1). The model had a mean accuracy of 0.919 (standard deviation [SD] = 0.028). The corresponding mean precision, recall, specificity, geometric mean, and F_β-Score were 0.521 (SD = 0.329), 0.051 (SD = 0.030), 0.994 (SD = 0.005), 0.337 (SD = 0.042), and 0.054 (SD = 0.029), respectively (Table 2).

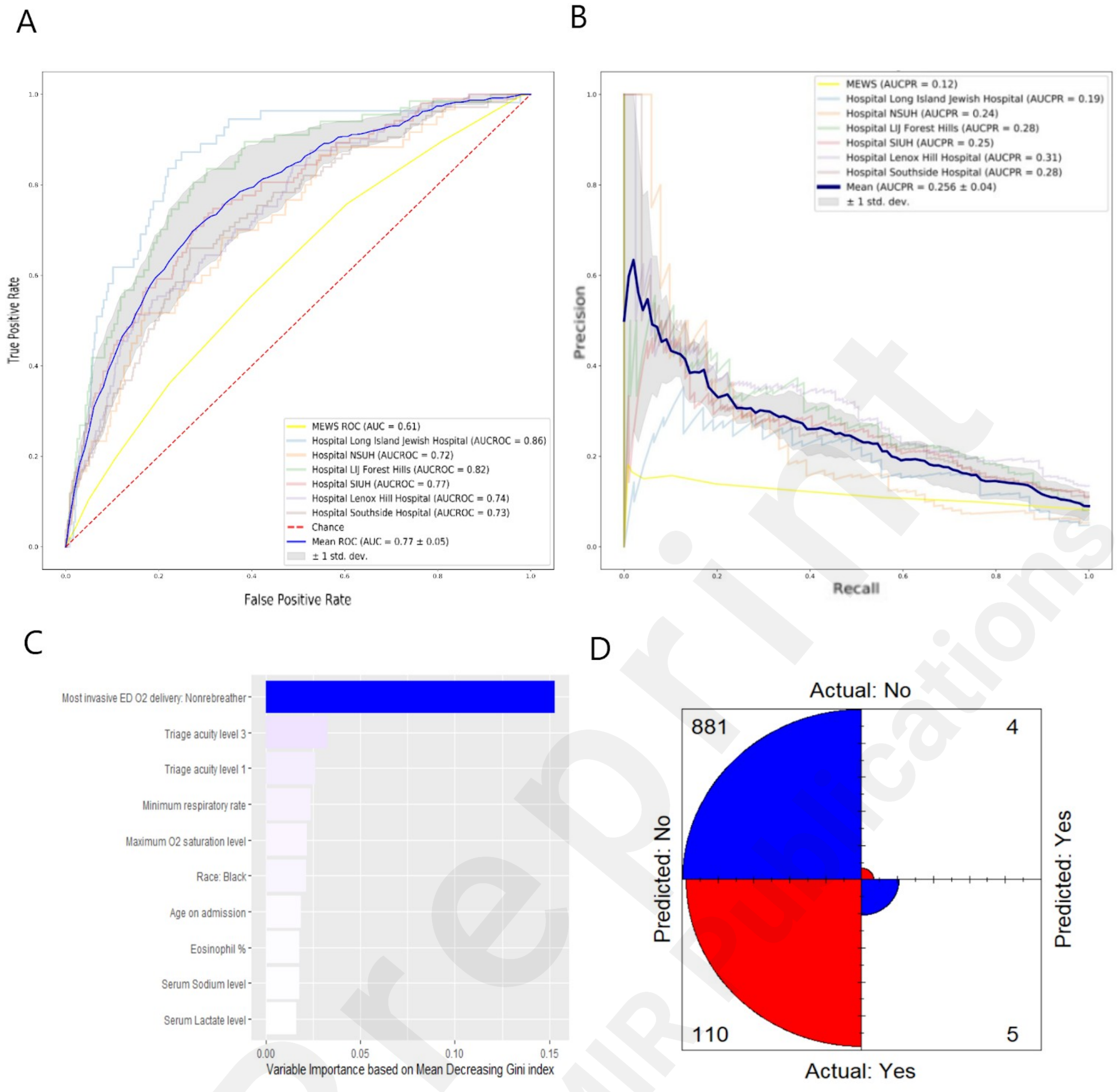


Figure 1. The XGBoost model for predicting respiratory failure within 48 hours. (*A*) ROC curve and (*B*) PR curve based on a cross hospital validation by leaving a hospital out as a testing set and the rest training set. Only hospitals with >1,000 COVID-19 patients were selected for testing sets. The mean ROC and PR curves are shown in dark blue and their corresponding ± SD are shown in gray. The MEWS score metrics are shown in light yellow. (*C*) Measurement of the 10 variables with the highest relative importance based on the amount they reduced the gini coefficient for the largest hospital testing set. (*D*) Confusion matrix visually represents the predicted values vs. actual prediction for the largest hospital testing set. AUC = area under the curve of ROC; AUCPR = area under the curve of precision-recall curve; MEWS = Modified Early Warning Score; PR = precision-recall; ROC = receiver operating characteristics; SD = standard deviation.

Table 2. Mean AUCROCs, AUCPRs, Accuracies, Precisions, Recalls, Specificities, Geometric Means and F_β-Score ($\beta = 4$) for models investigated.

Measure	XGBoost,	XGBoost +	LogisticReg,	MEWS

	mean (SD)	SMOTEENN (SD)	mean (SD)	
AUCROC	0.77 (0.05)	0.76 (0.03)	0.70 (0.05)	0.61
AUCPR	0.26 (0.04)	0.24 (0.06)	0.18 (0.06)	0.12
Accuracy	0.919 (0.028)	0.893 (0.016)	0.915 (0.027)	0.913
Precision	0.521 (0.329)	0.303 (0.089)	0.322 (0.375)	0.165
Recall	0.051 (0.030)	0.228 (0.095)	0.009 (0.013)	0.017
Specificity	0.994 (0.005)	0.955 (0.005)	0.998 (0.002)	0.992
Geometric mean	0.337 (0.042)	0.506 (0.063)	0.285 (0.051)	0.296
F_β-Score	0.054 (0.029)	0.226 (0.088)	0.010(0.014)	0.018

Definition of abbreviations: LogisticReg = Logistic Regression; MEWS = Modified Early Warning Score; SD = standard deviation.

Based on the XGBoost + SMOTEENN model, the mean AUCs of the ROC and PR curves were 0.76 ± 0.03 and 0.24 ± 0.06, respectively (Figure 2). The 10 most important variables, in order of decreasing importance, were: most invasive mode of oxygen delivery being a non-rebreather mask, ESI value of 3, male gender, white race, minimum respiratory rate, black race, ESI value of 2, most invasive mode of oxygen delivery being nasal cannula, ESI value of 1, and Hispanic ethnicity (Figure 2). The mean confusion matrix showed that most false predictions were false positives (those who were predicted to require intubation but were not intubated within 48 hours). False negatives (those who were predicted to not require intubation but were intubated within 48 hours) were the minority of predictions (Figure 2). While this model did not have the highest accuracy, it achieved the highest mean recall, geometric mean, and F_β-Score of 0.228 (SD = 0.095), 0.508 (SD = 0.063), and 0.226 (SD = 0.010), respectively. The corresponding mean accuracy, precision, and specificity were 0.893 (SD = 0.016), 0.303 (SD = 0.089), and 0.955 (SD = 0.005), respectively (Table 2).

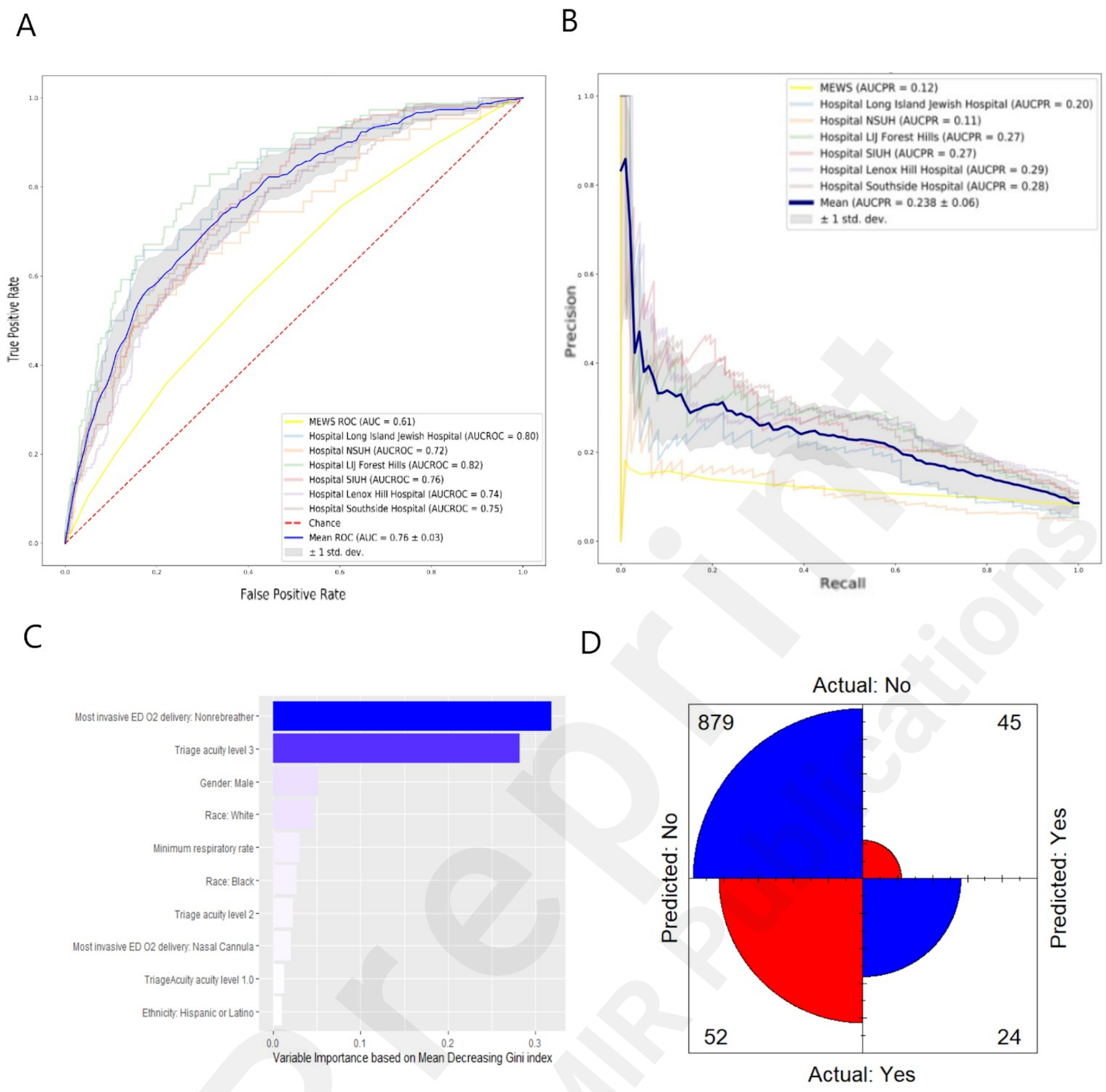


Figure 2. The XGBoost + SMOTEENN model for predicting respiratory failure within 48 hours. (*A*) ROC curve and (*B*) PR curve based on a cross hospital validation by leaving one hospital out as a testing set and the remaining constituting training set. Only hospitals with >1,000 COVID-19 patients were selected for testing sets. The mean ROC and PR curves are shown in dark blue and their corresponding ± SD are shown in gray. The MEWS score metrics are shown in light yellow. (*C*) The 10 variables with the highest relative importance measured by the amount the variable reduced the gini coefficient. (*D*) Mean confusion matrix visually represents the predicted values vs. actual prediction. AUC = area under the curve of ROC; AUCPR = area under the curve of precision-recall curve; MEWS = Modified Early Warning Score; PR = precision-recall; ROC = receiver operating characteristics; SD = standard deviation.

We also examined the performance of a Logistic Regression model. The mean AUCs of the ROC and PR curves were 0.70 ± 0.05 and 0.18 ± 0.06. Mean accuracy, precision, recall, specificity, geometric mean, and F$_\beta$-Score were 0.915 (SD = 0.027) ,0.322 (SD=0.375), 0.009 (SD = 0.013), 0.994 (SD = 0.005), 0.285 (SD = 0.051), and 0.010 (SD=0.014), respectively (Figure 3 and Table 2).

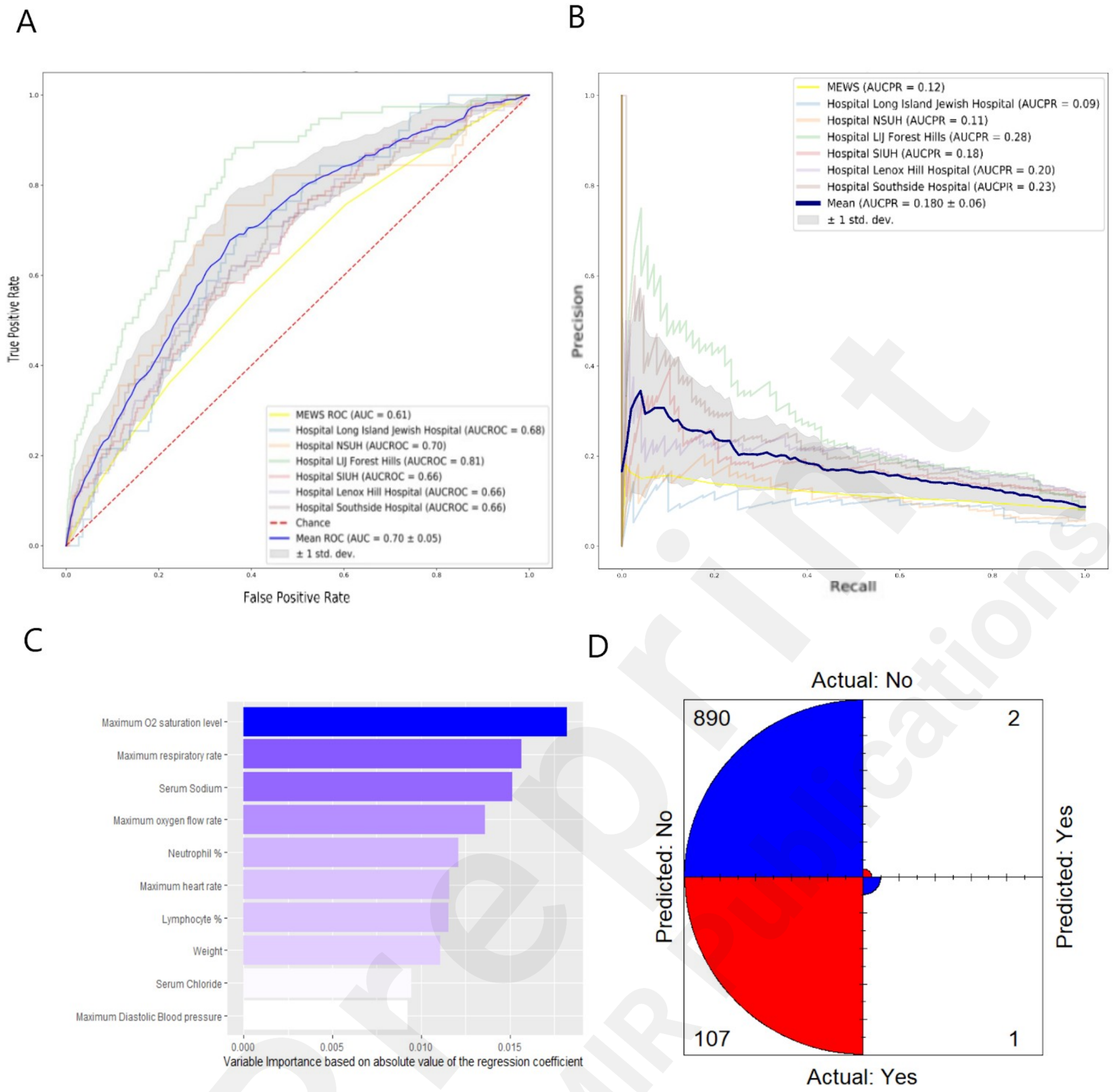


Figure 3. The Logistic regression model for predicting respiratory failure within 48 hours. (*A*) ROC curve and (*B*) PR curve based on a cross hospital validation by leaving a hospital out as a testing set and the rest training set. Only hospitals with >1000 COVID-19 patients were selected for testing sets. The mean ROC and PR curves are shown in dark blue and their corresponding ± SD are shown in gray. The MEWS score metrics are shown in light yellow. (*C*) Then 10 variables with the highest relative importance measured by the absolute value of the regression coefficient. (*D*) Mean confusion matrix visually represents the predicted values vs. actual prediction. AUC = area under the curve of ROC; AUCPR = area under the curve of precision-recall curve; MEWS = Modified Early Warning Score; PR = precision-recall; ROC = receiver operating characteristics; SD = standard deviation.

MEWS was used to compare ROC and PR curves. MEWS resulted in AUCs of the ROC and PR curves of 0.61 and 0.12, respectively (Figure 1, 2, and 3). For MEWS, accuracy, precision, recall, specificity, geometric mean, and F$_\beta$-Score were 0.913, 0.165, 0.017, 0.992, 0.296, and 0.018, respectively.

The calibration curves showed that all three models were well calibrated among all hospital folds. Although all three deviated from perfect calibration as the fraction of positives increased (Figure 3). The corresponding mean brier score for XGBoost, XGBoost + SMOTEENN, and Logistic Regression were 0.071 ± 0.019, 0.079 ± 0.016, and 0.077 ± 0.018 respectively (Figure 3).

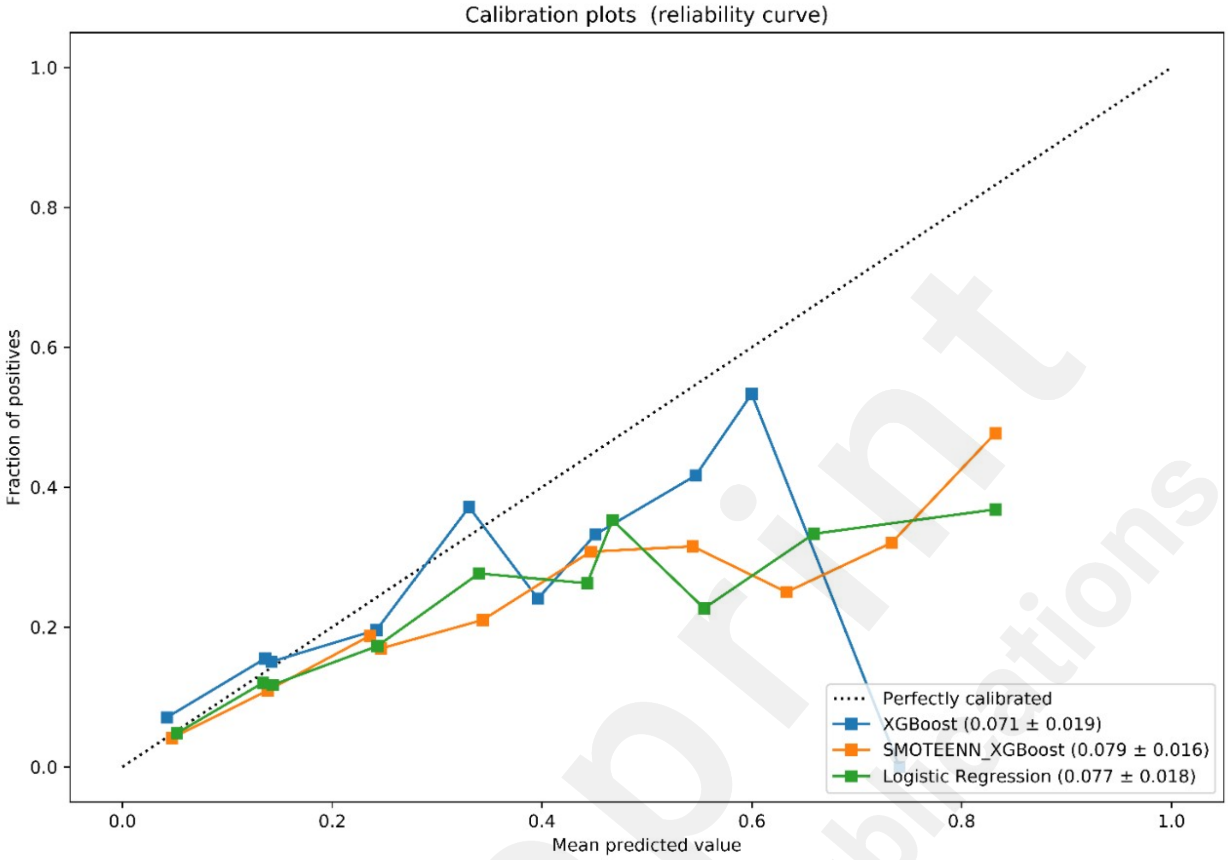


Figure 4. Calibration plots (reliability curve) of the XGBoost , XGBoost + SMOTEENN, and Logistic Regression for respiratory failure within 48 hours. Calibration is based on the precision probability (using predict_proba of python). For creating the plots, sklearn.calibration.CalibratedClassifierCV of python was used by inserting a fraction of positives and mean predicted values into 10 bins with increasing fraction of positive (respiratory failures) for each hospital fold . Mean Brier score ± SD across all hospitals tested corresponding to the model is shown in the figure legend in parentheses. SD: standard deviation.

Discussion

We presented three models for predicting early respiratory failure in patients given a diagnosis of COVID-19 and admitted to the hospital from the ED, two of which were based on XGBoost. One was tilted towards precision and specificity (XGBoost) and the other was tilted towards recall (XGBoost + SMOTEENN). These models are based on baseline characteristics, ED vital signs, and laboratory measurements. Using an automated tool to estimate the probability of respiratory failure could identify at-risk patients for earlier interventions (e.g., closer monitoring, critical care consultation, earlier discussions about goals of care) and improve patient outcomes.

We evaluated three machine learning models: XGBoost, XGBoost + SMOTEENN, and Logistic Regression [38–40]. XGBoost is widely used due to its high efficiency and predictability, and it has been used to predict healthcare outcomes in patients with [41,42] and without [43–45] COVID-19. In our study, XGBoost was the most accurate prediction model, with an accuracy of 0.919 (SD = 0.028) and precision of 0.521 (SD=0.329) (Figure 1), similar to the findings of another study that examined combined outcomes [46]. However, what is different in our model is that it achieves cross-hospital validation. Such accuracy showcases the ability of the model to separate intubations from non-intubations within the 48-hour window of interest. Such a model would be useful for physicians as it more accurately and consistently identifies patients at high risk for intubation.

We also constructed an XGBoost + SMOTEENN model. SMOTEENN was used to improve the sensitivity of our prediction, because our data is imbalanced (i.e., only ~8% of our COVID-19 cohort were intubated), while keeping deviation from accuracy and calibration of the model to a minimum. Compared to XGBoost, the XGBoost + SMOTEENN model had lower accuracy and precision, but greater recall (or sensitivity) (0.228 (SD=0.095)) (Figure 2). This higher sensitivity can identify more patients who require IMV, suggesting that this model may be more suitable for broad or automated screening of patients.

We also examined the performance of a Logistic Regression model to determine whether a compact, linear model could accurately predict patient risk (Figure 3 and Table 2). Model performance was inferior to the XGBoost model. This supports earlier reports that machine learning techniques outperform classic models of logistic regression in their ability to predict many prognostic and health outcomes [47–49]. Finally, we compared the performance of our predictive machine learning models to the widely used Modified Early Warning Score (MEWS) [36]. MEWS was inferior to all three models described above in most of the measures examined.

Using the most important variables for our models, we identified clinically relevant measures that can best inform clinical decision making (Figure 1, 2). The XGBoost model was accurate and precise, as reflected in the low number of false positives of the model predictions (Figure 1). A more sensitive alternative to this model would be the XGBoost + SMOTEENN model, which had fewer false negatives than XGBoost (Figure 2). Both models share important predictors, such as information about the mode of oxygen delivery, triage acuity, demographic information, and respiratory rate. However, XGBoost (the more accurate model with higher precision) adds serum lactate, sodium, and eosinophil percentage to the top 10 most important variables. as well. This shows when precision is important, measures such as lactate can rule out the most severe cases by becoming strong predictors. Among hospitals in Northwell health, certain hospitals such as Long Island Jewish (which is one of the largest in terms of number of Covid19 patients) have a high drop in their predictive ability when logistic regression is used. When Long Island Jewish is being validated, the 0.86 AUCROC of the XGBoost model drops to 0.68 for Logistic Regression. This could partially be due to the nature of the outcome predicted (choice of ventilation from hospital staff), where one would expect different hospitals to possibly exhibit higher variability, not only to patient demographics, but also to hospital staff therapy choices.

Variable importance metrics revealed that the linear logistic regression models use laboratory variables primarily, whereas nonlinear XGBoost based models prioritize clinical, demographic and variables that better capture hospital specific behavior (oxygen delivery types prior to intubation) and increase the robustness of the model. However, we need to validate whether providing these variables along with the probability of respiratory failure would decrease the rate of identifying at-risk patients. Further prospective studies and randomized clinical trials are needed for this validation.

When examining calibration of the models (Figure 4), we found that all models were well calibrated, yet as the fraction of positive cases increased, calibration suffered. This suggests, that if a specific population of patients has greater likelihood of intubation (for example, those age > 70, or with specific comorbid conditions), the model would need to be retrained to increase its accuracy and calibration.

Our study has several limitations. We extracted data on intubation timing from our EHR, which may have minor inaccuracies. While a consistent temporal inaccuracy could create bias in underestimating/overestimating the intubation rate, we believe that these small inaccuracies are overcome by the average calculated from our large number of cases. Another limitation is that we relied on data from a multicenter, single health system for both implementation and validation. Thus, we were unable to externally validate the models in other health systems and hospitals with different protocols, which might affect the model’s performance. Also, because the study is retrospective, we can only suggest associations and correlations rather than identify the main contributors that lead to intubation and mechanical ventilation. Furthermore, the numerical missing variables were imputed with weighted k-nearest neighbors. Thus, the conclusions made from these variables assume uniformity in patient data based on those missing values. In the case of non-uniformity, the order of variable importance might change. Also, some clinical variables included in the model may appear obvious correlates of the clinical decision for intubation within 48 hours (e.g., having non-rebreather oxygen flow as the most invasive form of ventilation). However, the association of all included variables is not deterministic: only 453 of 2633 patients on non-rebreather oxygen flow in the ED were intubated within 48 hours. Also, given that these variables are available to clinicians and part of their decision-making, we included them in our model. Finally, we used supervised learning on a homogenous database. While we used cross hospital validation and retrospectively validated our learning method, external generalizability of these learning methods to other health systems requires validation in prospective studies and randomized trials. Such high-quality evidence could provide more clues on clinical and economic impacts, as well as measures to improve them.

COVID-19 has presented an extremely challenging clinical and public emergency worldwide, especially in the New York City metropolitan area. As public health measures attempt to mitigate this disaster by slowing the spread and alleviating the heavy burden placed on health care systems, clinicians must make important decisions quickly and hospital administrators must manage resources and personnel. Furthermore, as predicted by many models [50–52] , we are in the midst of a second second wave of infection. Our models could inform clinical care by offering complementary performance characteristics (one model with superior recall, the other with greater precision) and supporting clinical decision-making as we tackle this unprecedented public health crisis.

Acknowledgements

We acknowledge and honor all of our Northwell team members who consistently put themselves in harm’s way during the COVID-19 pandemic. We dedicate this article to them, as their vital contribution to knowledge about COVID-19 and sacrifices on the behalf of patients made it possible.
SB, DPB, and TPZ conceptualized and designed the study. SB, JSH, and TPZ had full access to all data in the study and are responsible for the integrity of the data. SB and JSH performed data extraction and cleaning. MB, PW, TM, and JSH contributed to many discussions during manuscript development. SB, DPB, and TPZ contributed to drafts of the manuscript. SB trained and validated the models. SB and TPZ designed and created the figures. DBP and TPZ critically reviewed the paper, and PW and TPZ obtained funding. The Northwell COVID-19 Research Consortium prioritized this manuscript, organized meetings between contributing authors, and provided support in finalizing the manuscript for submission. All named authors read and approved the final submitted manuscript.

Sources of support:

This work was supported by R24AG064191 from the National Institute on Aging and R01LM012836 from the National Library of Medicine, and R35GM118337 from the National Institute for General Medical Sciences, all of the National Institutes of Health**.**

Conflict of Interest: The authors confirm that there are no conflicts of interest.

References

1. WHO Director-General’s opening remarks at the media briefing on COVID-19 - 11 March 2020. . cited 2020 May 8. Available from: https://www.who.int/dg/speeches/detail/who-director-general-s-opening-remarks-at-the-media-briefing-on-covid-19---11-march-2020

2. The New York Times. Coronavirus in the U.S.: Latest Map and Case Count. The New York Times [Internet]. 2020 Mar 3 . cited 2020 May 8; Available from: https://www.nytimes.com/interactive/2020/us/coronavirus-us-cases.html

3. Richardson S, Hirsch JS, Narasimhan M, Crawford JM, McGinn T, Davidson KW, the Northwell COVID-19 Research Consortium, Barnaby DP, Becker LB, Chelico JD, Cohen SL, Cookingham J, Coppa K, Diefenbach MA, Dominello AJ, Duer-Hefele J, Falzon L, Gitlin J, Hajizadeh N, Harvin TG, Hirschwerk DA, Kim EJ, Kozel ZM, Marrast LM, Mogavero JN, Osorio GA, Qiu M, Zanos TP. Presenting Characteristics, Comorbidities, and Outcomes Among 5700 Patients Hospitalized With COVID-19 in the New York City Area. JAMA. 2020 May 26; 323(20):2052–2059. PMCID: PMC7177629

4. Myers LC, Parodi SM, Escobar GJ, Liu VX. Characteristics of Hospitalized Adults With COVID-19 in an Integrated Health Care System in California. JAMA. 2020 ;323(21):2195. PMCID: PMC7182961

5. Tsertsvadze T, Ezugbaia M, Endeladze M, Ratiani L, Javakhishvili N, Mumladze L, Khotchava M, Janashia M, Zviadadze D, Gopodze L, Gokhelashvili A, Metchurchtlishvili R, Abutidze A, Chkhartishvili N. Characteristics and outcomes of hospitalized adult COVID-19 patients in Georgia. Available from: http://dx.doi.org/10.1101/2020.10.23.20218255

6. Ruan Q, Yang K, Wang W, Jiang L, Song J. Clinical predictors of mortality due to COVID-19 based on an analysis of data of 150 patients from Wuhan, China. Intensive Care Medicine. 2020; 46(5):846–848. PMCID: PMC7080116

7. Ruan Q, Yang K, Wang W, Jiang L, Song J. Correction to: Clinical predictors of mortality due to COVID-19 based on an analysis of data of 150 patients from Wuhan, China. Intensive Care Medicine. 2020; 46(6):1294–1297.

8. Goyal P, Choi JJ, Pinheiro LC, Schenck EJ, Chen R, Jabri A, Satlin MJ, Campion TR Jr, Nahid M, Ringel JB, Hoffman KL, Alshak MN, Li HA, Wehmeyer GT, Rajan M, Reshetnyak E, Hupert N, Horn EM, Martinez FJ, Gulick RM, Safford MM. Clinical Characteristics of Covid-19 in New York City. N Engl J Med. 2020 Jun 11; 382(24):2372–2374. PMCID: PMC7182018

9. Levitan R. Opinion: The Infection That’s Silently Killing Coronavirus Patients. The New York Times [Internet]. 2020 Apr 20 . cited 2020 May 8; Available from: https://www.nytimes.com/2020/04/20/opinion/sunday/coronavirus-testing-pneumonia.html

10. Tobin MJ. Basing Respiratory Management of Coronavirus on Physiological Principles. Am J Respir Crit Care Med. 2020 Apr 13; 201(11):1319–1320. PMID: 32281885

11. Kangelaris KN, Ware LB, Wang CY, Janz DR, Zhuo H, Matthay MA, Calfee CS. Timing of intubation and clinical outcomes in adults with acute respiratory distress syndrome. Crit Care Med. 2016 Jan; 44(1):120–129. PMCID: PMC4774861

12. Serin SO, Karaoren G, Esquinas AM. Delayed admission to ICU in acute respiratory failure: Critical time for critical conditions. The American journal of emergency medicine. 2017:1571–1572. PMID: 28502761

13. Renaud B, Santin A, Coma E, Camus N, Van Pelt D, Hayon J, Gurgui M, Roupie E, Hervé J, Fine MJ, Brun-Buisson C, Labarère J. Association between timing of intensive care unit admission and outcomes

for emergency department patients with community-acquired pneumonia. Crit Care Med. 2009 Nov; 37(11):2867–2874. PMID: 19770748

14. Liu V, Kipnis P, Rizk NW, Escobar GJ. Adverse outcomes associated with delayed intensive care unit transfers in an integrated healthcare system. J Hosp Med. 2012 Mar; 7(3):224–230. PMID: 22038879

15. Churpek MM, Carey KA, Dela Merced N, Prister J, Brofman J, Edelson DP. Validation of Early Warning Scores at Two Long-Term Acute Care Hospitals. Crit Care Med. 2019 Sep 24; 47(12): e962–e965. PMID: 31567342

16. Churpek MM, Yuen TC, Winslow C, Meltzer DO, Kattan MW, Edelson DP. Multicenter Comparison of Machine Learning Methods and Conventional Regression for Predicting Clinical Deterioration on the Wards. Crit Care Med. 2016 Feb; 44(2):368–374. PMCID: PMC4736499

17. Dziadzko MA, Novotny PJ, Sloan J, Gajic O, Herasevich V, Mirhaji P, Wu Y, Gong MN. Multicenter derivation and validation of an early warning score for acute respiratory failure or death in the hospital. Crit Care. 2018 Oct 30; 22(1):286. PMCID: PMC6206729

18. Yu S, Leung S, Heo M, Soto GJ, Shah RT, Gunda S, Gong MN. Comparison of risk prediction scoring systems for ward patients: a retrospective nested case-control study. Crit Care. 2014 Jun 26;18(3): R132. PMCID: PMC4227284

19. Subbe CP, Slater A, Menon D, Gemmell L. Validation of physiological scoring systems in the accident and emergency department. Emerg Med J. 2006; 23:841–845. PMID: 17057134

20. Debnath S, Barnaby DP, Coppa K, Makhnevich A, Kim EJ, Chatterjee S, Tóth V, Levy TJ, Paradis MD, Cohen SL, Hirsch JS, Zanos TP, Northwell COVID-19 Research Consortium. Machine learning to assist clinical decision-making during the COVID-19 pandemic. Bioelectron Med. 2020 Jul 10; 6:14. PMCID: PMC7347420

21. Ferrari D, Milic J, Tonelli R, Ghinelli F, Meschiari M, Volpi S, Faltoni M, Franceschi G, Iadisernia V, Yaacoub D, Ciusa G, Bacca E, Rogati C, Tutone M, Burastero G, Raimondi A, Menozzi M, Franceschini E, Cuomo G, Corradi L, Orlando G, Santoro A, Digaetano M, Puzzolante C, Carli F, Borghi V, Bedini A, Fantini R, Tabbì L, Castaniere I, Busani S, Clini E, Girardis M, Sarti M, Cossarizza A, Mussini C, Mandreoli F, Missier P, Guaraldi G. Machine learning in predicting respiratory failure in patients with COVID-19 pneumonia—Challenges, strengths, and opportunities in a global health emergency. PLOS ONE. 2020; 15(11): e0239172. PMCID: PMC7660476

22. Assaf D, Gutman Y 'ara, Neuman Y, Segal G, Amit S, Gefen-Halevi S, Shilo N, Epstein A, Mor-Cohen R, Biber A, Rahav G, Levy I, Tirosh A. Utilization of machine-learning models to accurately predict the risk for critical COVID-19. Intern Emerg Med. 2020 Nov; 15(8):1435–1443. PMCID: PMC7433773

23. Haimovich AD, Ravindra NG, Stoytchev S, Young HP, Wilson FP, van Dijk D, Schulz WL, Taylor RA. Development and Validation of the Quick COVID-19 Severity Index: A Prognostic Tool for Early Clinical Decompensation. Ann Emerg Med. 2020 Oct; 76(4):442–453. PMCID: PMC7373004

24. Batista GEAPA, Gustavo E A P, Prati RC, Monard MC. A study of the behavior of several methods for balancing machine learning training data. 2004: 20–29. Available from: http://dx.doi.org/10.1145/1007730.1007735

25. Yu H-F, Huang F-L, Lin C-J. Dual coordinate descent methods for logistic regression and maximum entropy models. Machine Learning. 2011; 85(1-2):41–75.

26. Chen T, Guestrin C. XGBoost: A Scalable Tree Boosting System. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA. 2016: 785–794.

27. Lemaître G, Nogueira F, Aridas CK. Imbalanced-learn: A Python Toolbox to Tackle the Curse of Imbalanced Datasets in Machine Learning. J Mach Learn Res. 2017; 18(17):1–5.

28. More A. Survey of resampling techniques for improving classification performance in unbalanced datasets. arXiv [stat.AP]. 2016. Available from: http://arxiv.org/abs/1608.06048

29. Troyanskaya O, Cantor M, Sherlock G, Brown P, Hastie T, Tibshirani R, Botstein D, Altman RB. Missing value estimation methods for DNA microarrays. Bioinformatics. 2001;17(6):520–525. PMID: 11395428

30. Harris SL, Harris DM. Sequential Logic Design. Digital Design and Computer Architecture. 2016:108–171.

31. Buitinck L, Louppe G, Blondel M, Pedregosa F, Mueller A, Grisel O, Niculae V, Prettenhofer P, Gramfort A, Grobler J, Layton R, VanderPlas J, Joly A, Holt B, Varoquaux G. API design for machine learning software: experiences from the scikit-learn project. ECML PKDD Workshop: Languages for Data Mining and Machine Learning. 2013:108–122.

32. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J, Passos A, Cournapeau D, Brucher M, Perrot M, Duchesnay E. Scikit-learn: Machine Learning in Python, Imputation transformer for completing missing values. . cited 2019 Dec 26. Available from: https://scikit-learn.org/stable/modules/generated/sklearn.impute.SimpleImputer.html

33. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J, Passos A, Cournapeau D, Brucher M, Perrot M, Duchesnay E. Scikit-learn: Machine Learning in Python, Imputation for completing missing values using k-Nearest Neighbors. cited 2020 May 6. Available from: https://scikit-learn.org/stable/modules/generated/sklearn.impute.KNNImputer.html

34. Gerry S, Birks J, Bonnici T, Watkinson PJ, Kirtley S, Collins GS. Early warning scores for detecting deterioration in adult hospital patients: a systematic review protocol. BMJ Open. 2017 Dec 3;7(12): e019268. PMCID: PMC5736035

35. Bilben B, Grandal L, Søvik S. National Early Warning Score (NEWS) as an emergency department predictor of disease severity and 90-day survival in the acutely dyspneic patient – a prospective observational study. Scandinavian Journal of Trauma, Resuscitation and Emergency Medicine. 2016; 2:24–80. PMID: 27250249

36. Subbe CP, Kruger M, Rutherford P, Gemmel L. Validation of a modified Early Warning Score in medical admissions. QJM. 2001 Oct;94(10):521–526. PMID: 11588210

37. Hripcsak G. Writing Arden Syntax medical logic modules. Computers in Biology and Medicine. 1994;24(5):331–363. PMID: 7705066

38. Sheppard C. Tree-based Machine Learning Algorithms: Decision Trees, Random Forests, and Boosting. Createspace Independent Publishing Platform; 2017.

39. Mani I, Zhang I. kNN approach to unbalanced data distributions: a case study involving information extraction. Proceedings of workshop on learning from imbalanced datasets [Internet]. site.uottawa.ca; 2003. Available from: https://www.site.uottawa.ca/~nat/Workshop2003/jzhang.pdf

40. Yu H-F, Huang F-L, Lin C-J. Dual coordinate descent methods for logistic regression and maximum entropy models. Machine Learning. 2011; 85(1-2):41–75.

41. Yan L, Zhang H-T, Goncalves J, Xiao Y, Wang M, Guo Y, Sun C, Tang X, Jing L, Zhang M, Huang X, Xiao Y, Cao H, Chen Y, Ren T, Wang F, Xiao Y, Huang S, Tan X, Huang N, Jiao B, Cheng C, Zhang Y, Luo A, Mombaerts L, Jin J, Cao Z, Li S, Xu H, Yuan Y. An interpretable mortality prediction model for

COVID-19 patients. Nature Machine Intelligence. 2020 May 1; 2(5):283–288.

42. Kumar A, Gupta PK, Srivastava A. A review of modern technologies for tackling COVID-19 pandemic. Diabetes Metab Syndr. Elsevier; 2020 May 7; 14(4):569–573. PMCID: PMC7204706

43. Xu Z, Wang Z. A Risk Prediction Model for Type 2 Diabetes Based on Weighted Feature Selection of Random Forest and XGBoost Ensemble Classifier. 2019 Eleventh International Conference on Advanced Computational Intelligence (ICACI) [Internet]. 2019. Available from: http://dx.doi.org/10.1109/icaci.2019.8778622

44. Sharma A, Willem J M. Improving Diagnosis of Depression With XGBOOST Machine Learning Model and a Large Biomarkers Dutch Dataset (n = 11,081). Frontiers in Big Data. 2020; 3:15.

45. Zabihi M, Kiranyaz S, Gabbouj M. Sepsis Prediction in Intensive Care Unit Using Ensemble of XGboost Models. 2019 Computing in Cardiology Conference (CinC) [Internet]. 2019. Available from: http://dx.doi.org/10.22489/cinc.2019.238

46. 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, China Medical Treatment Expert Group for COVID-19. Development and Validation of a Clinical Risk Score to Predict the Occurrence of Critical Illness in Hospitalized Patients With COVID-19. JAMA Intern Med. 2020 May 12;180(8):1–9. PMCID: PMC7218676

47. Kourou K, Exarchos TP, Exarchos KP, Karamouzis MV, Fotiadis DI. Machine learning applications in cancer prognosis and prediction. Computational and Structural Biotechnology Journal. 2015; 13:8–17.

48. Liu NT, Salinas J. Machine Learning for Predicting Outcomes in Trauma. Shock. 2017 Nov; 48(5):504–510. PMID: 28498299

49. Ferroni P, Zanzotto F, Riondino S, Scarpato N, Guadagni F, Roselli M. Breast Cancer Prognosis Using a Machine Learning Approach. Cancers. 2019:328. Available from: http://dx.doi.org/10.3390/cancers11030328

50. Xu S, Li Y. Beware of the second wave of COVID-19. The Lancet. 2020:1321–1322. PMCID: PMC7194658

51. Bikbov B, Bikbov A. Communication on COVID-19 to community – measures to prevent a second wave of epidemic. Int J Health Sci (Qassim). 2020 May;14(3):1–3. PMCID: PMC7269624

52. Strzelecki A. The second worldwide wave of interest in coronavirus since the COVID-19 outbreaks in South Korea, Italy and Iran: A Google Trends study. Brain Behav Immun. 2020 Apr 18;88:950–951. PMCID: PMC7165085

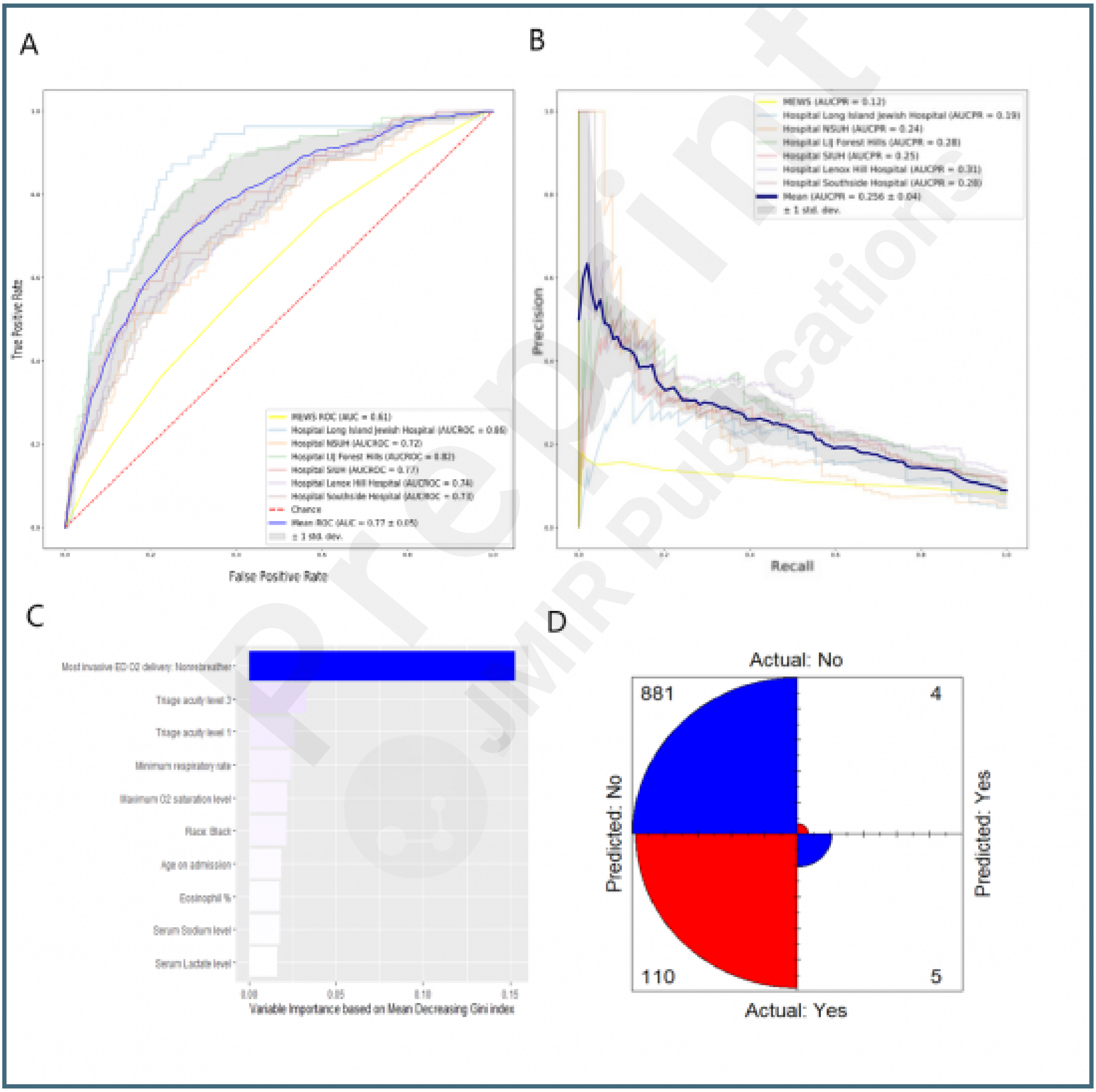
Abbreviation list

ROC: Receiver operating characteristic
PR: Precision-Recall curve
ED: Emergency department
COVID-19: coronavirus disease 2019
ICU: Intensive care unit
IMV: Invasive mechanical ventilation
EHR: Electronic health record
MEWS: Modified early warning score
AUC: Area under the curve
AUCPR: Area under the PR curve
LDH: Lactate dehydrogenase
BMI: Body mass index
SD: Standard deviation
IQR: Interquartile range
SBP: Systolic blood pressure
DBP: Diastolic blood pressure
COPD: Chronic obstructive pulmonary disease
HR: Heart rate
pO_2: Partial pressure of oxygen
RR: Respiratory rate
LogisticReg: Logistic regression
SMOTE: Synthetic Minority Oversampling
ENN: Edited Nearest Neighbours
SMOTEENN: Over-sampling using SMOTE and cleaning using ENN

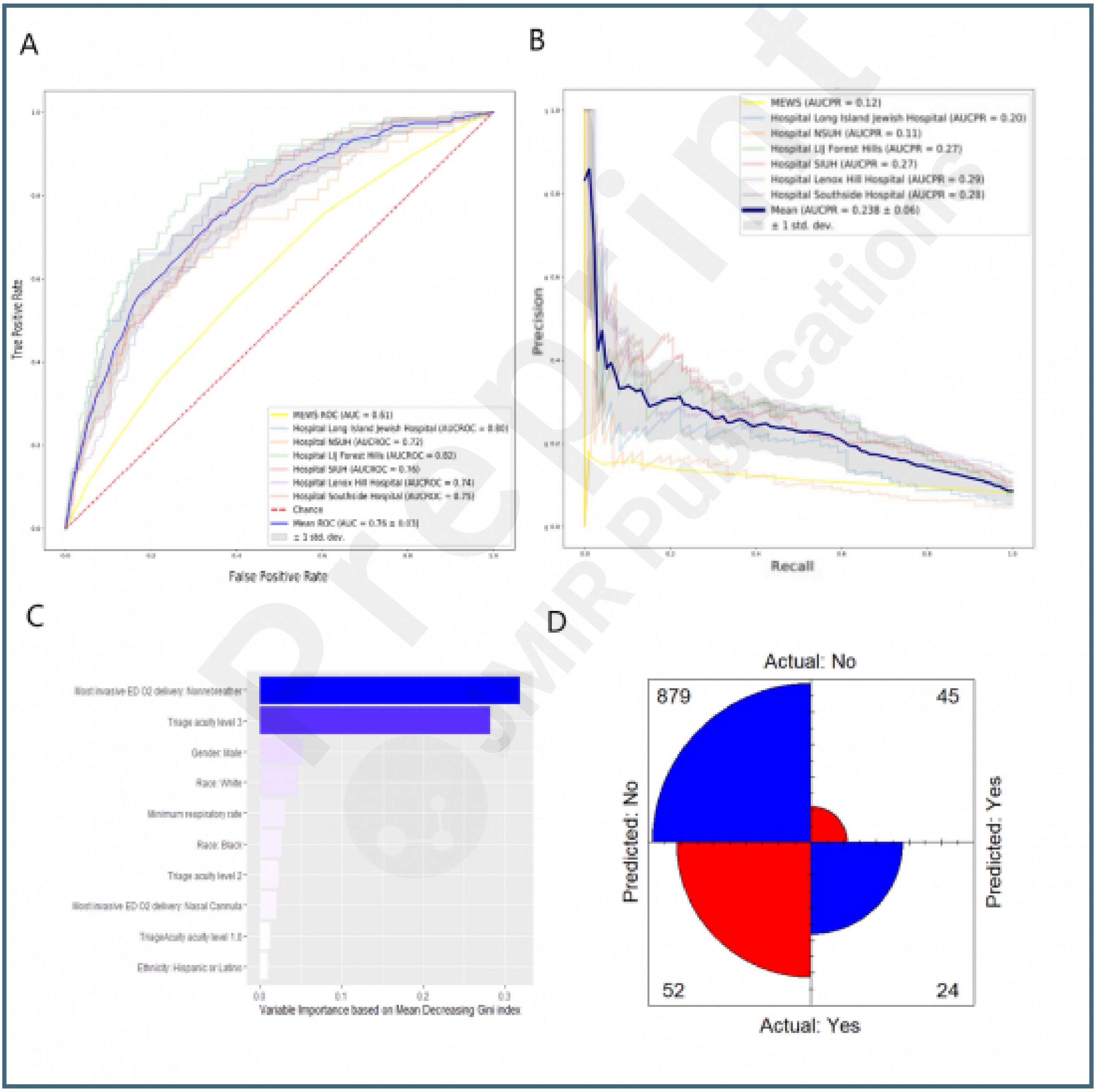
Supplementary Files

Figures

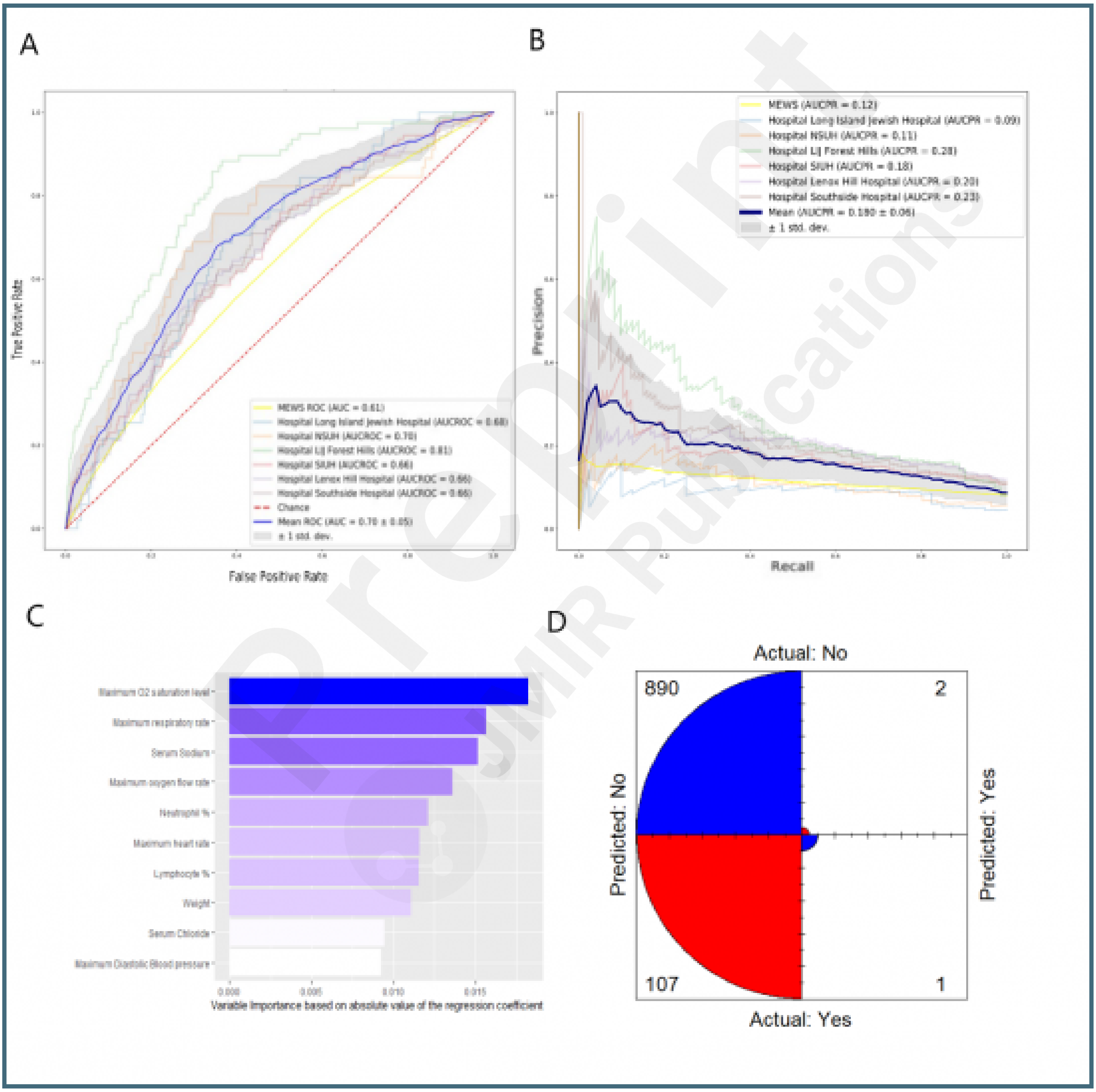
The XGBoost model for predicting respiratory failure within 48 hours. (A) ROC curve and (B) PR curve based on a cross hospital validation by leaving a hospital out as a testing set and the rest training set. Only hospitals with >1,000 COVID-19 patients were selected for testing sets. The mean ROC and PR curves are shown in dark blue and their corresponding ± SD are shown in gray. The MEWS score metrics are shown in light yellow. (C) Measurement of the 10 variables with the highest relative importance based on the amount they reduced the gini coefficient for the largest hospital testing set. (D) Confusion matrix visually represents the predicted values vs. actual prediction for the largest hospital testing set. AUC = area under the curve of ROC; AUCPR = area under the curve of precision-recall curve; MEWS = Modified Early Warning Score; PR = precision-recall; ROC = receiver operating characteristics; SD = standard deviation.



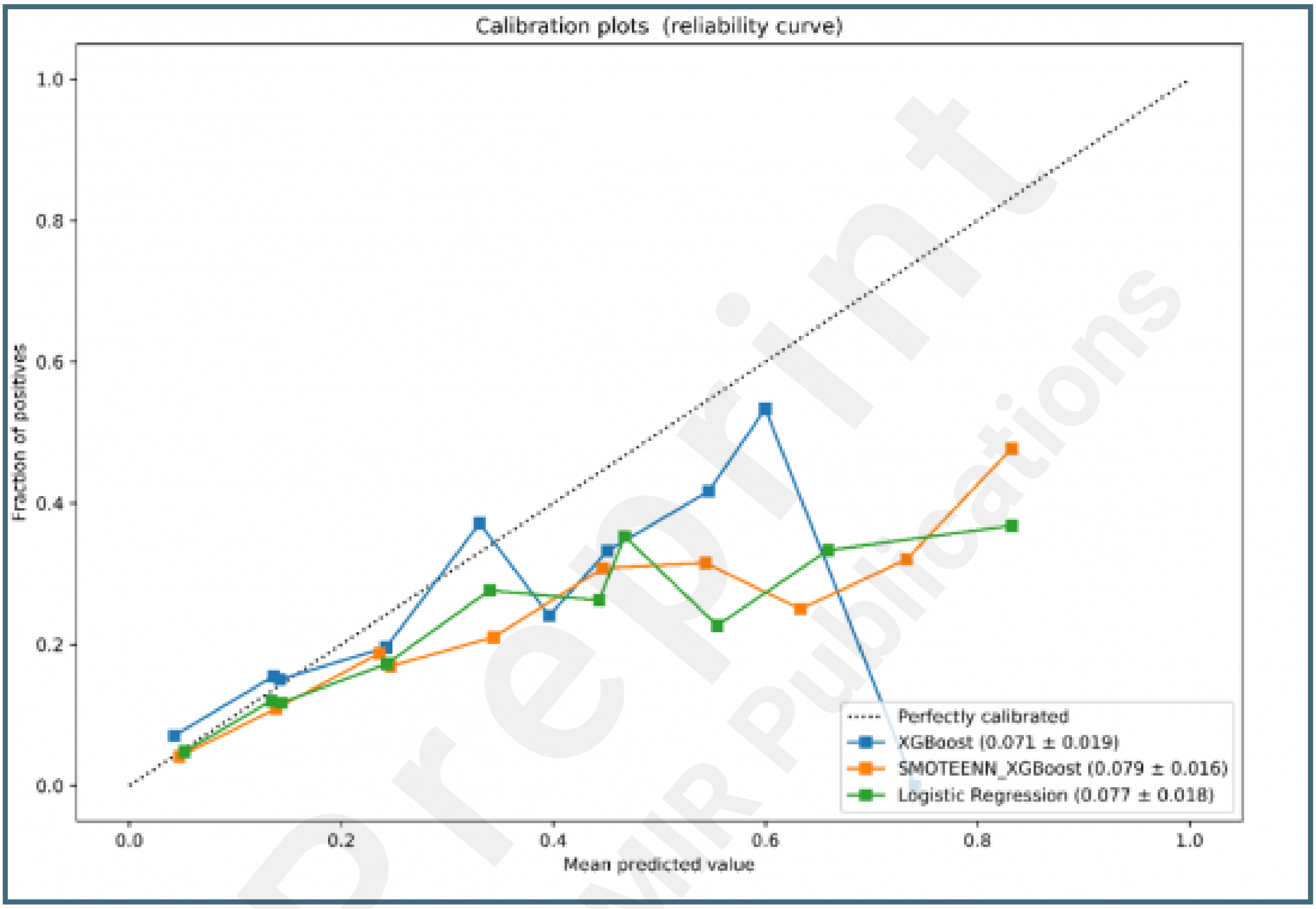
The XGBoost + SMOTEENN model for predicting respiratory failure within 48 hours. (A) ROC curve and (B) PR curve based on a cross hospital validation by leaving one hospital out as a testing set and the remaining constituting training set. Only hospitals with >1,000 COVID-19 patients were selected for testing sets. The mean ROC and PR curves are shown in dark blue and their corresponding ± SD are shown in gray. The MEWS score metrics are shown in light yellow. (C) The 10 variables with the highest relative importance measured by the amount the variable reduced the gini coefficient. (D) Mean confusion matrix visually represents the predicted values vs. actual prediction. AUC = area under the curve of ROC; AUCPR = area under the curve of precision-recall curve; MEWS = Modified Early Warning Score; PR = precision-recall; ROC = receiver operating characteristics; SD = standard deviation.



The Logistic regression model for predicting respiratory failure within 48 hours. (A) ROC curve and (B) PR curve based on a cross hospital validation by leaving a hospital out as a testing set and the rest training set. Only hospitals with >1,000 COVID-19 patients were selected for testing sets. The mean ROC and PR curves are shown in dark blue and their corresponding ± SD are shown in gray. The MEWS score metrics are shown in light yellow. (C) Then 10 variables with the highest relative importance measured by the absolute value of the regression coefficient. (D) Mean confusion matrix visually represents the predicted values vs. actual prediction. AUC = area under the curve of ROC; AUCPR = area under the curve of precision-recall curve; MEWS = Modified Early Warning Score; PR = precision-recall; ROC = receiver operating characteristics; SD = standard deviation.



Calibration plots (reliability curve) of the XGBoost , XGBoost + SMOTEENN, and Logistic Regression for respiratory failure within 48 hours. Calibration is based on the precision probability (using predict_proba of python). For creating the plots, sklearn.calibration.CalibratedClassifierCV of python was used by inserting a fraction of positives and mean predicted values into 10 bins with increasing fraction of positive (respiratory failures) for each hospital fold . Mean Brier score ± SD across all hospitals tested corresponding to the model is shown in the figure legend in parentheses. SD: standard deviation.



Multimedia Appendixes

Table. Definitions of Accuracy, Precision, Recall, Specificity, Geometric Means and F?-Score.
URL: https://asset.jmir.pub/assets/fcbe7234df9cfc8334871eb902e8d019.docx

Table. Modified Early Warning Score calculation based on vital sign measurements.
URL: https://asset.jmir.pub/assets/dbec4887335b5642fb871c13e0b2892f.docx