Research and Applications

Development and validation of predictive models for COVID-19 outcomes in a safety-net hospital population

Boran Hao[1,2], Yang Hu[1,2], Shahabeddin Sotudian[1,3], Zahra Zad[1,3], William G. Adams[4], Sabrina A. Assoumou[5], Heather Hsu[4], Rebecca G. Mishuris[5], and Ioannis C. Paschalidis[1,2,3,6]

[1]Center for Information and Systems Engineering, Boston University, Boston, Massachusetts, USA, [2]Department of Electrical and Computer Engineering, Boston University, Boston, Massachusetts, USA, [3]Division of Systems Engineering, Boston University, Boston, Massachusetts, USA, [4]Department of Pediatrics, Boston Medical Center and Boston University School of Medicine, Boston, Massachusetts, USA, [5]Department of Medicine, Boston Medical Center and Boston University School of Medicine, Boston, Massachusetts, USA, and [6]Department of Biomedical Engineering, Boston University, Boston, Massachusetts, USA

Corresponding Author: Ioannis C. Paschalidis, Division of Systems Engineering, Department of Electrical and Computer Engineering, Department of Biomedical Engineering, and Faculty of Computing & Data Sciences, Boston University, 8 Saint Mary's St., Boston, MA 02215, USA; yannisp@bu.edu http://sites.bu.edu/paschalidis

ABSTRACT

Objective: To develop predictive models of coronavirus disease 2019 (COVID-19) outcomes, elucidate the influence of socioeconomic factors, and assess algorithmic racial fairness using a racially diverse patient population with high social needs.

Materials and Methods: Data included 7,102 patients with positive (RT-PCR) severe acute respiratory syndrome coronavirus 2 test at a safety-net system in Massachusetts. Linear and nonlinear classification methods were applied. A score based on a recurrent neural network and a transformer architecture was developed to capture the dynamic evolution of vital signs. Combined with patient characteristics, clinical variables, and hospital occupancy measures, this dynamic vital score was used to train predictive models.

Results: Hospitalizations can be predicted with an area under the receiver-operating characteristic curve (AUC) of 92% using symptoms, hospital occupancy, and patient characteristics, including social determinants of health. Parsimonious models to predict intensive care, mechanical ventilation, and mortality that used the most recent labs and vitals exhibited AUCs of 92.7%, 91.2%, and 94%, respectively. Early predictive models, using labs and vital signs closer to admission had AUCs of 81.1%, 84.9%, and 92%, respectively.

Discussion: The most accurate models exhibit racial bias, being more likely to falsely predict that Black patients will be hospitalized. Models that are only based on the dynamic vital score exhibited accuracies close to the best parsimonious models, although the latter also used laboratories.

Conclusions: This large study demonstrates that COVID-19 severity may accurately be predicted using a score that accounts for the dynamic evolution of vital signs. Further, race, social determinants of health, and hospital occupancy play an important role.

Key words: COVID-19, AI, predictive modeling, racial bias, social determinants of health

INTRODUCTION

Coronavirus disease 2019 (COVID-19) has affected more than 450 million people globally. Although about 65% of the US population has been vaccinated against severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2),[1] rates of immunization have been uneven especially among different racial/ethnic groups and between rural versus urban communities.[2] Limited vaccination rates and the emergence of new variants suggests that COVID-19 will remain a concern for health systems worldwide.[3]

Making predictions about disease severity is important in clinical triage, resource allocation, staffing, and overall planning, within a hospital system, and at the state/country scale. Artificial Intelligence (AI) methods have been used to that end,[4] including the prediction of patient outcomes for COVID-19.[5–11] However, these studies used data from relatively few patients (the largest used 2,500[6]) and a limited collection of pre-existing conditions, laboratories, and in-hospital data.

More importantly, no predictive models of hospitalization, disease severity, and mortality have been developed using data from a safety-net hospital caring for a large percentage of racially/ethnically diverse patients, including many lower-income individuals with pressing needs associated with social determinants of health (SDOH). In addition, no models have leveraged SDOH for patients receiving clinical care. While work exploring disparities has considered aggregate nationwide data in the United States[12] and Brazil,[13] there is a need for more detailed analysis[14] and a concern that, if not properly adjusted, models may perpetuate biases.[15] The present study includes a large percentage of Black and Hispanic patients and person-level information on SDOH, enabling a characterization of specific race/ethnic and SDOH variables that influence the predictive models.

An additional characteristic of the current study is the availability of rich information on the daily/hourly evolution of vitals for hospitalized patients. Most of the previously published predictive models were static—they considered a snapshot of the patient's condition and made a forward prediction. In the current work, we leveraged neural networks with long short-term memory cells[16] and a transformer[17] encoder to build a score of vitals that captures their dynamic evolution. Models based just on this score perform surprisingly well compared to more complex models that also use a host of laboratories. In addition, access to hospital occupancy data reveals how they may influence care decisions.

MATERIALS AND METHODS

Data description

We de-identified data for all 7,102 patients with a positive reverse transcription polymerase chain reaction (RT-PCR) SARS-CoV-2 test at the Boston Medical Center (BMC) between January 1 and December 31, 2020. As a tertiary care academic medical center, BMC is the largest safety-net hospital in New England, providing care for about 30% of Boston residents.[18] Features extracted included demographics, SDOH variables, depression status, travel-contact information, vital signs, radiological findings, past medical history, symptoms, medications, laboratory tests, hospital occupancy, hospitalization course, admission to the Intensive Care Unit (ICU), mechanical ventilation, and mortality. SDOH variables were based on answers to the THRIVE survey administered at BMC, which identifies social needs in 8 domains: housing, food, medication, transportation, utilities, childcare, employment, and education.[19] We also

used self-reported race and ethnicity in the electronic health records and hospital occupancy, which was measured by the daily bed usage percentage for surgeries, COVID, and non-COVID patients. The Supplementary Material includes additional details. The study was approved by the BMC Institutional Review Board.

Preprocessing and variable selection

We developed predictive models for the following outcomes: (1) hospitalization, (2) ICU care, (3) mechanical ventilation, and (4) mortality. For each patient, we built a profile containing all outcome labels and extracted features. Instead of using computer vision techniques to extract information from radiology images,[20] we used natural language processing (NLP) to extract radiology findings from text (see Supplementary Material). We applied "one-hot" encoding to represent categorical features as 0 and 1. We retained variables for which we had values from at least 350 patients, and we imputed the missing values in a continuous-valued feature using the mean of its nonmissing values. All features were standardized to zero mean and unit standard deviation.

For the hospitalization model, we used the admission date as reference date, and the earliest positive SARS-CoV-2 test as the reference date for non-admitted patients. Features were extracted according to the reference dates. We utilized all features except laboratory results, medication, and radiological findings, which were typically not available for non-hospitalized patients. The features we used for predicting hospitalization include pre-existing conditions and SDOH information from the patient's hospital record, symptoms, and observed vital signs. All these features would have been readily available to physicians in either the emergency room or the outpatient clinics making these decisions.

For admitted patients, the closest records before their reference date were extracted and we only included records that were within 48 h before the reference date. For non-admitted patients, we only included records that were within 48 h before or after the reference date. For ICU, mechanical ventilation and mortality prediction models, we only considered admitted patients. In addition to the features utilized in the hospitalization model, laboratory results and radiological findings were used, and we excluded symptoms since severe COVID-19 patients were less likely to describe their symptoms. The earliest and latest vitals and lab results to be used for ICU/Intubation/Mortality models depend on the timeline settings for models, which will be introduced in "Timeline Strategy" section. For patients with the identified outcomes (ICU care, mechanical ventilation, death), the date of the outcome was used as the reference date. For patients with absent outcomes, a random date during their hospitalization was used as their reference date. In general, all input features for predicting the various in-hospital outcomes would have been readily available to physicians in advance of the predicted outcome.

Timeline strategy

We introduced a timeline strategy to capture the dynamic evolution of vital signs, labs, and radiological findings for predicting ICU, mechanical ventilation, and mortality. Given the reference date t and a desired "drop time" τ_0, we first eliminated all features during the interval $[t - \tau_0, \ t]$. Then, we defined k consecutive time windows with length τ each, tracing back from $t - \tau_0$. The mean (for continuous features) or maximum (for categorical features) of all feature records in the ith time window $[t - \tau_0 - i\tau, \ t - \tau_0 - (i - 1)\tau]$ was computed and defined as "feature $- i\tau$." We used the sequence "feature $-$

1τ,"…,"feature − $k\tau$" as a feature timeline to train the models. We did not implement timelines for the hospitalization model, because laboratory and radiology findings were not used and vital sign records were sparse.

Classification methods

We applied linear and nonlinear classifiers to predict outcomes. Linear methods included logistic regression (LR) and support vector machines (SVM).[21] Nonlinear methods included XGBoost[22] and Random Forest (RF).[23] We introduced regularizations to prevent the influence of outliers in data.[24,25] Furthermore, we used LSTM-Transformer neural networks to compute a score capturing the dynamic evolution of vitals over the timeline.

We applied statistical feature selection (SFS), removing variables with high p-value. We removed one from each pair of features with absolute correlation coefficient >0.8. We further implemented ℓ1-regularized LR recursive feature elimination (RFE). Features retained from RFE were used to derive a parsimonious LR model (see Supplementary Material for details).

Model evaluation

We evaluated model performance using 2 metrics: area under the curve (AUC) of the receiver-operating characteristic (ROC) and Weighted-F1 score. ROC plots the recall (or sensitivity) against the false positive rate, and AUC can be interpreted as the probability that a randomly chosen sample from the positive class will score higher than a randomly chosen sample from the negative class. The F1 score is the harmonic mean of recall and precision. The weighted-F1 score is calculated by weighting the F1 score of each class by the number of samples in that class. Values for both metrics are between 0 to 1 and a higher value implies a better model.

We split patients into a training (80%) and test set (20%). We trained the models on the training set and evaluated them on the test set. We repeated this procedure 5 times, each with a different random split. The average and standard deviation on the test set over the 5 random splits are reported. For each split we further applied 5-fold cross-validation on the training set to find the best hyperparameters of each model; therefore, the test set is completely independent and kept separate from the training process. We performed external validation to assess the generalizability of our hospitalization models. We trained hospitalization models using all BMC samples and evaluated their performance on data from Mass General Brigham used in our earlier work.[6] We did not attempt external validation for other models because they rely on clinical variables and it was not possible to match those across the 2 data sets.

We compared our models with the NEWS2[26] score for predicting deterioration and the sepsis score qSOFA.[27] These are computed from vital signs, so we compared them with our LSTM-Transformer vital score. In addition, we trained the "BMC protocol," which is a classifier using a group of labs and vital signs chosen for evaluating COVID-19 severity by BMC physicians (see Supplementary Material).

RESULTS

Among 7,102 patients, 19.5% were admitted. Among the hospitalized, 23.3% required ICU care, 13.8% received mechanical ventilation, and 9.65% died. The mean age of all patients was 47.9 and 35.1% were Black. Representative statistics are in Supplementary Table S1 (see Supplementary Material).

Hospitalization models

Prediction models

The hospitalization model used the entire data set, labeling patients as hospitalized (class 1) or non-hospitalized (class 0). About 126 variables for each patient were retained after preprocessing. The average of the obtained metrics over 5 random splits is reported in Table 1. We compared the performance of linear (ie, best performing SVM and LR) and nonlinear (ie, XGBoost and RF) methods using all 126 variables. After SFS, 70 variables were retained and RFE retained 20 variables. The latter "parsimonious" model was enhanced by adding 2 hospital utilization variables, while controlling for additional relevant variables. Specifically, for each patient we added "Total Non-COVID Percentage" and "Total COVID Percentage," indicating the ratio of the number of patients treated for non-COVID diseases and COVID, respectively, over the total number of BMC beds, computed at the patient's reference time. This resulted into parsimonious models with 22 variables.

The parsimonious models performed almost as well as the models with all 126 features. Table 1 also reports the composition of an ℓ2-regularized LR model. Larger values of the variables with positive (respectively, negative) coefficient increase (respectively, decrease) the likelihood of hospitalization. For instance, the likelihood of hospitalization decreases with increased hospital occupancy. Two SDOH variables (Food insecurity and need for Transportation) were observed to increase hospitalization likelihood.

External validation

We trained the model with the 20 variables retained after RFE on all BMC patients and evaluated its performance on patients from 5 hospitals in the Mass General Brigham system used in earlier work[6] (Table 1).

Analysis focused on racial bias

We retained 249 Black and 251 White patients for testing and trained a model (with the 22 features of the parsimonious model) on the rest of the patients. Table 1 presents the performance of this model on the 2 cohorts. We used a treatment equality[28] definition to evaluate the fairness of the hospitalization treatment, which requires the ratio of the false positive rate (FPR) over the false negative rate (FNR) to be equal among the 2 cohorts. This ratio is 73.6% higher for black patients than for whites. Note that we are controlling for the most important variables associated with a hospitalization, hence, this bias is due to unmeasured factors not used by the model, or possibly from missing values of variables the model uses that may more severely affect one of the cohorts.

To resolve this racial bias, we modified the prediction threshold of the LR model (to which the predicted likelihood is compared). The default value for this threshold is 0.5. We selected 2 different thresholds, 1 for black patients and 1 for white patients, seeking to equalize the FPR/FNR ratio while keeping the FNR relatively low (around 0.25). Table 1 reports these thresholds and the resulting metrics.

ICU models

The ICU prediction results are in Table 2. We first trained one immediate model (0-drop), using the features in the past 36 h to predict need for immediate ICU care. For vitals we used a $k = 6$, $\tau = 6$ h timeline, while for laboratory and radiology findings we only used one $\tau = 36$ h window, since most laboratory data and imaging were taken at most once a day. After combining the vitals into the

Table 1. Hospitalization prediction models

Algorithm	AUC	F1-weighted
	Models using all 126 features	
LR-L1	86.5% (0.9%)	87.0% (0.5%)
SVM-L1	86.3% (0.9%)	86.8% (0.8%)
XGBoost	93.1% (0.6%)	90.1% (0.8%)
RF	92.4% (0.5%)	90.0% (0.7%)
	Models using 70 statistically selected features	
LR-L1	86.1% (1.0%)	87.3% (0.7%)
SVM-L1	86.0% (1.0%)	87.1% (0.6%)
XGBoost	92.5% (0.5%)	90.2% (0.7%)
RF	92.2% (0.5%)	89.9% (0.7%)
	Parsimonious Model using 22 features	
LR-L1	83.5% (1.7%)	85.3% (0.7%)
SVM-L1	83.4% (1.8%)	84.7% (1.0%)
XGBoost	92.0% (0.6%)	89.6% (0.6%)
RF	91.3% (0.6%)	89.4% (0.5%)

Variables for the Parsimonious Model

Variable	Coef	OR	OR 95% CI		Y1-mean	Y0-mean	Y-corr	*p*-value
Intercept	−0.7	–	–	–	–	–	–	–
Dyspnea	1.28	3.98	3.3	4.8	0.36	0.08	0.32	1.60E-47
PMH-Transplantation	0.72	3.36	2.06	5.5	0.04	0.01	0.09	1.32E-06
Vomiting	0.52	1.86	1.26	2.74	0.05	0.02	0.07	1.88E-03
PMH-CHF	0.42	1.64	1.17	2.3	0.1	0.02	0.16	4.50E-03
PMH-Hepatitis	0.55	1.54	1.14	2.09	0.07	0.04	0.05	5.05E-03
SpO2	−0.83	0.68	0.65	0.72	96.09	98	−0.33	8.06E-56
SDOH-Food	0.28	1.4	0.87	2.24	0.03	0.02	0.02	1.61E-01
PMH-CKD	0.39	1.35	1.02	1.79	0.14	0.04	0.16	3.65E-02
PMH-Stroke	0.35	1.32	0.9	1.94	0.06	0.02	0.1	1.50E-01
PMH-COPD	−0.16	0.79	0.56	1.11	0.06	0.03	0.06	1.75E-01
Abdominal Pain	0.23	1.26	0.9	1.77	0.06	0.03	0.07	1.75E-01
PMH-Arrhythmia	0.28	1.25	0.9	1.75	0.08	0.03	0.12	1.88E-01
Contact-COVID-Travel	0.23	1.24	1.06	1.45	0.37	0.28	0.08	6.24E-03
Respiratory Rate	0.45	1.18	1.14	1.22	20.07	18.07	0.28	4.87E-25
Diarrhea	0.14	1.1	0.81	1.5	0.07	0.04	0.06	5.39E-01
SDOH-Transportation	0.23	1.06	0.58	1.95	0.02	0.01	0.03	8.43E-01
BMI	0.08	1.04	1.02	1.06	30.01	29.12	0.1	1.65E-05
Diastolic BP	−0.15	0.97	0.96	0.98	75.28	77.63	−0.13	2.56E-12
Age	0.52	1.03	1.03	1.04	57.31	45.62	0.26	1.13E-44
PMH-Cirrhosis	0.1	1.02	0.52	2.01	0.02	0.01	0.03	9.52E-01
Total non-COVID percentage	−0.55	0.991	0.989	0.994	69.05	73.71	−0.08	1.11E-10
Total COVID percentage	−0.59	0.992	0.99	0.994	40.72	38.26	0.04	2.43E-11

Performance evaluation on MGH Brigham data

Algorithm	AUC	F1-weighted
	Models using all 20 features	
LR-L1	81.30%	77.10%
SVM-L1	82.20%	77.80%
XGBoost	83.00%	77.80%
RF	79.10%	75.30%

Racial disparity in the LR-L2 parsimonious hospitalization model

	Threshold	FPR	FNR	FPR/FNR	AUC	F1-weighted
Black patients	0.5	0.199	0.288	0.691	81.90%	78.93%
White patients	0.5	0.148	0.372	0.398	80.44%	82.39%

Table 1. Continued

Controlling for racial disparity in the LR-L2 parsimonious hospitalization model

	Threshold	FPR	FNR	FPR/FNR	F1-weighted
Black patients	0.45	0.242	0.237	1.021	77.36%
White patients	0.41	0.239	0.233	1.025	78.71%

Note: The values inside the parentheses denote the standard deviation of the corresponding metric. SVM-L1 and LR-L1 refer to the ℓ_1-norm regularized SVM and LR models. We report the composition of an ℓ_2-norm regularized LR model, including the coefficient of each variable (Coef), the correlation of the variable with the outcome (Y-corr), the mean of the variable (Y1-mean) in the hospitalized, and the mean of the variable (Y0-mean) in the non-hospitalized. For each variable, we also report the corresponding *p*-value, the odds ratio (OR), and its 95% confidence interval (CI).

SpO2: oxygen saturation; BP: blood pressure; BMI: body mass index; PMH: past medical history; CKD: chronic kidney disease; COPD: chronic obstructive pulmonary disease; CHF: congestive heart failure; SDOH: social determinants of health; Total non-COVID percentage: (Total number of non-COVID patients at the hospital/Total number of beds)×100; Total COVID percentage: (Total number of COVID patients at the hospital/Total number of beds)×100.

LSTM-Transformer score, we selected at most 10 features using SFS and RFE (reported in Table 2) and trained a parsimonious model. The parsimonious model yielded an average AUC of 92.7%, which is close to the best full model AUC of 94.8%. Using only NEWS2 or qSOFA scores as the feature yielded AUC of 84.3% and 71.7%, respectively, lower than the AUC of 90.9% using the LSTM-Transformer vital score as the only feature.

We further trained 2 extreme models to predict if a patient would need ICU care after 12 h (12-h drop model) and 24 h (24-h drop model). By changing the drop time in the timeline to 12 and 24 h, respectively, the model is not using any information for the patient in the 12/24 h before ICU admission. The parsimonious models maintained a high AUC of 86.5% and 81.1%, respectively, which match and exceed the corresponding best nonlinear full models with AUC of 86.6% and 79.9%, respectively. While for these early predictions NEWS2 and qSOFA-based models performed poorly, the LSTM-Transformer score remained a strong predictor. Apparently, for immediate predictions all models did relatively well, whereas for longer term predictions the LSTM-Transformer score and other models including it show significant advantage.

Mechanical ventilation models

The mechanical ventilation prediction results are in Table 3. As with the ICU models, we trained one immediate model (0-drop), using the past 36 h features to predict if a patient needs to be intubated immediately. For vitals we used a $k = 6$, $\tau = 6$ h timeline, while for laboratory and radiology findings we only used one $\tau = 36$ h window. After combining vitals into the LSTM-Transformer score, we selected 10 features using RFE and trained a parsimonious model. The top features are reported in Table 3. The parsimonious model obtained an average AUC of 91.2%, close to the AUC of the best full model (93.8%). Using only NEWS2 or qSOFA scores yields an AUC of 66.0% and 63.1%, respectively, lower than the AUC of 90.0% obtained by using just the LSTM-Transformer vital score.

We further trained 2 extreme models to predict if a patient would need intubation after 12 h (12-h drop model) and 24 h (24-h drop model); the corresponding parsimonious models have AUC of 90.3% and 84.9%, respectively. NEWS2- and qSOFA-based models do considerably worse in these advance predictions.

Mortality models

Due to the relatively longer mean time gap between hospitalization and death, we built different timelines for the mortality models. The first mortality model only uses features within 3 d after admission (adm-based model), and $k = 3$, $\tau = 24$ h are applied in this timeline.

Consequently, we can predict a patient's mortality at the very early stage of hospitalization. Another model uses a drop time of 24 h prior to death (24-h drop model), using $k = 7$ and $\tau = 48$ h for the timeline. For both settings, the LSTM-Transformer vital score is used in parsimonious models. Performance and top features are reported in Table 4.

For adm-based models, the best full model achieved 91.4% AUC, while the parsimonious model using LR with only 13 features did better (AUC of 92.0%). The AUC of qSOFA and NEWS2 models did not exceed 69%, and the LSTM-Transformer score yielded a model with AUC of 84.3%. For the 24-h drop model, the best nonlinear model achieved 96.2% AUC, and the parsimonious model using 13 features achieved an AUC of 94.0%. When the outcome draws near, the advantage of the LSTM-Transformer score over the NEWS2 score remains significant.

DISCUSSION

The best AUCs achieved by the 4 models are between 93% and 96%, indicating strong predictive power. Strong predictions are achieved with relatively few features used by parsimonious models. These models use no more than 22 features each for hospitalization and mortality prediction, and no more than 10 features each for ICU and ventilation prediction, yielding similar (or better) performance with an AUC differential of −2.6% to +1.2% compared to the best models. This indicates the possibility of implementing simple, actionable predictive models to aid triage, staffing, and resource planning. The models produced outperformed related models in the literature (eg, the ventilation model outperforms an earlier model with a 74% AUC[29]).

Patients' vital signs were the most important factors for ICU, ventilation, and mortality prediction. These vital signs imply the severity of the disease and the potential need for cardiorespiratory resuscitation. Most of the prior studies use vital signs as "static" independent predictive variables.[6–9] In this study, we used an LSTM + Transformer encoder deep neural network to develop a single score combining all vitals and capturing their dynamic evolution over time. Models for ICU and ventilation (short-term and longer term) predictions using just the LSTM-Transformer vital score have an AUC within 1.2–4.9% from the corresponding parsimonious models which also use other clinical variables; essentially, for these models, vital sign trends alone suffice! For mortality predictions, the LSTM-Transformer score is the top variable but other clinical variables significantly enhance performance.

Long-term predictions are more challenging than short-term: ICU, ventilation, and mortality predictions deteriorate as we move

Table 2. ICU prediction models

	0-drop		12 h-drop		24 h-drop	
	AUC	F1-weighted	AUC	F1-weighted	AUC	F1-weighted
Best full models before SFS	XGBoost		XGBoost		XGBoost	
	94.8% (1.2%)	89.2% (1.8%)	86.6% (1.2%)	82.4% (0.9%)	78.3% (2.0%)	76.0% (2.3%)
Best full models after SFS	XGBoost (112 features)		XGBoost (104 features)		XGBoost (95 features)	
	94.3% (1.4%)	88.0% (2.3%)	86.3% (1.8%)	82.2% (1.2%)	79.9% (3.2%)	76.8% (3.4%)
Parsimonious models LR-L1-LSTM-Transformer	92.7% (1.5%)	86.7% (1.3%)	86.5% (1.6%)	81.6% (2.1%)	81.1% (2.5%)	79.4% (0.8%)
BMC-Protocol LR-L1	89.0% (2.2%)	86.1% (2.0%)	73.0% (2.0%)	73.7% (1.0%)	67.4% (1.7%)	70.8% (1.7%)
BMC-Protocol XGBoost	94.8% (0.8%)	89.1% (1.7%)	86.2% (1.0%)	83.2% (1.0%)	77.1% (2.5%)	74.0% (1.4%)
NEWS2 score	84.3% (2.0%)	83.5% (1.8%)	48.6% (2.5%)	72.2% (1.4%)	46.5% (2.3%)	70.4% (1.6%)
qSOFA score	71.7% (3.0%)	79.3% (1.7%)	54.3% (2.1%)	69.1% (1.8%)	52.1% (2.4%)	68.1% (2.0%)
LSTM-Transformer score	90.9% (2.0%)	85.2% (1.8%)	84.5% (2.0%)	81.3% (1.2%)	76.8% (3.9%)	73.6% (1.5%)

Variable	Coef	Y1-mean	Y0-mean	*p*-value	Y-corr
0-drop parsimonious model 10 features					
LSTM-Transformer vital score	3.42	3.44	−4.70	5.63E-163	0.82
Intercept	−2.44	–	–	–	–
Glucose	0.52	182.20	129.24	5.52E-13	0.26
PMH-DVT	0.46	0.09	0.04	2.12E-02	0.08
LDH	0.45	496.37	356.83	2.19E-32	0.24
Total COVID percentage	−0.41	40.76	42.16	3.41E-01	−0.03
PMH-HLD	0.31	0.32	0.24	2.94E-02	0.08
CO_2	−0.24	22.62	24.13	2.53E-07	−0.19
NRBC Percentage	0.23	0.23	0.22	4.25E-03	0.01
Lung Opacity on X-Ray	0.2	0.68	0.29	2.35E-33	0.33
CRP	0.1	123.53	70.81	1.87E-28	0.30
12 h-drop parsimonious model 10 features					
LSTM-Transformer vital score	1.81	0.83	−1.18	1.79E-74	0.50
Intercept	−0.88	–	–	–	–
Total COVID percentage	−0.33	40.45	41.97	3.16E-01	−0.03
PMH-Arrhythmia	0.22	0.12	0.07	2.71E-02	0.08
LDH	0.22	431.34	358.87	1.49E-30	0.19
BUN	0.2	21.32	18.41	5.47E-20	0.09
Age	0.16	60.09	56.47	2.20E-03	0.09
CRP	0.14	100.25	72.09	4.97E-33	0.19
Ferritin	0.12	1758.64	990.79	4.78E-29	0.15
PMH-CKD	0.11	0.20	0.12	1.14E-02	0.09
PMH-CAD	0.11	0.09	0.04	1.19E-02	0.09

Note: For each full model, we only report results from the algorithm with the highest AUC out of LR, SVM, XGBoost, and RF. We present the LR coefficients of each variable (Coef), the correlation of the variable with the outcome (Y-corr), the *p*-value, the mean of the variable (Y1-mean) in the ICU patients, and the mean of the variable (Y0-mean) in the non-ICU patients.

LDH: lactate dehydrogenase; BUN: blood urea nitrogen; NRBC: nucleated red blood cell; CKD: chronic kidney disease; PMH: past medical history; CAD: coronary artery disease; DVT: deep vein thrombosis; HLD: hypersensitivity lung disease; Total COVID percentage: (Total number of COVID patients at the hospital/Total number of beds)×100.

further from the time of the outcome. While most models do relatively well for short-term predictions, the parsimonious models which include the LSTM-Transformer score increase their advantage to baseline models (eg, NEWS2 and qSOFA) when longer term predictions are sought. Specifically, the AUC differential between the parsimonious model and the best of the NEWS2 and qSOFA-based models increases from 8.4–25.2% for short-term predictions to 17.3–32.2% for longer term predictions. Incidentally, the protocol used at BMC fares better and is closer to the parsimonious model for both short and long-term predictions.

Some of the variables included in the ICU prediction model have previously been identified in the literature. Patient age and past medical history like renal disease (CKD), cardiac disease (CAD) have extensively been described as factors influencing disease severity.[30]

Laboratory data such as CRP and ferritin[31] are acute phase reactants and have also previously been associated with COVID-19 disease severity. The large and diverse population used in our work strongly supported these findings, and our interpretable LR model coefficients further numerically show their relative importance. The mortality model also includes laboratory data that have also previously been identified in the literature as being associated with disease severity such as CRP, ferritin, and LDH.[6,32] Since mortality prediction models use multiple time windows for labs as well, the most informative period of a certain lab is further revealed.

Analyzing data from a safety net hospital with a high proportion of Black patients and information on SDOH needs, gave us an opportunity to assess the effect of racial bias and socioeconomic variables. We elected to consider potential racial bias only between Black

Table 3. Mechanical ventilation prediction models

	0-drop		12 h-drop		24 h-drop	
	AUC	F1-weighted	AUC	F1-weighted	AUC	F1-weighted
Best full models before SFS	XGBoost		RF		RF	
	93.6% (0.9%)	91.6% (0.8%)	90.6% (1.2%)	87.8% (1.0%)	86.3% (1.1%)	85.0% (2.1%)
Best full models after SFS	XGBoost (90 features)		XGBoost (104 features)		XGBoost (107 features)	
	93.8% (1.2%)	91.8% (1.5%)	90.3% (1.4%)	88.8% (0.7%)	86.1% (1.1%)	85.9% (1.4%)
Parsimonious models	91.2% (2.0%)	90.7% (0.9%)	90.3% (1.4%)	87.6% (1.0%)	84.9% (1.5%)	84.7% (0.5%)
LR-L1-LSTM-Transformer						
BMC-Protocol	82.3% (0.9%)	85.0% (1.7%)	63.4% (3.1%)	79.9% (0.6%)	56.7% (1.5%)	79.9% (0.7%)
LR-L1						
BMC-Protocol XGBoost	90.0% (1.0%)	88.7% (0.9%)	88.1% (1.1%)	88.3% (0.3%)	84.0% (1.1%)	80.3% (0.8%)
NEWS2 score	66.0% (6.5%)	84.2% (1.3%)	65.7% (2.8%)	80.0% (0.0%)	67.6% (2.8%)	80.0% (0.0%)
qSOFA score	63.1% (5.1%)	80.0% (0.4%)	52.3% (2.0%)	80.0% (0.0%)	52.3% (2.0%)	80.0% (0.0%)
LSTM-Transformer score	90.0% (2.3%)	88.4% (1.6%)	85.9% (2.8%)	86.3% (1.2%)	80.0% (2.5%)	79.9% (0.2%)

Variable	Coef	Y1-mean	Y0-mean	*p*-value	Y-corr
0-drop parsimonious model 10 features					
LSTM-Transformer vital score	2.26	2.25	−3.44	4.97E-104	0.68
Intercept	−2.15	–	–	–	–
Immature Granulocytes 0 Percentage	−0.63	0.26	0.50	1.79E-08	−0.17
Gender (male)	0.43	0.68	0.53	3.17E-03	0.10
Venous PH	−0.42	7.32	7.35	6.15E-06	−0.11
Absolute Lymphocytes	−0.4	1.35	1.47	7.22E-07	−0.06
Red cell Distribution Width	0.3	14.54	14.25	1.14E-03	0.06
Glucose	0.3	166.23	130.88	5.32E-21	0.18
Total Elective Surgery	−0.3	6.68	9.65	3.05E-03	−0.08
CRP	0.21	118.04	71.76	1.06E-29	0.22
Age	0.19	59.76	56.92	4.72E-02	0.05
12 h-drop parsimonious model 10 features					
LSTM-Transformer vital score	1.54	2.42	0.18	5.05E-62	0.52
Intercept	−1.41	–	–	–	–
CRP	0.6	92.16	72.83	3.32E-29	0.11
Immature Granulocytes 0 Percentage	−0.55	0.09	0.44	2.76E-18	−0.25
Hemoglobin	0.47	12.45	12.34	1.96E-12	0.02
Red Cell Distribution Width	0.41	14.38	14.29	2.86E-14	0.02
Total Elective Surgery percentage	−0.4	6.61	9.75	3.56E-03	−0.08
Calcium	−0.37	8.82	8.87	1.52E-10	−0.04
CO_2	−0.36	23.65	24.04	1.51E-12	−0.05
Age	0.31	59.76	56.92	4.72E-02	0.05
Gender (male)	0.26	0.68	0.53	3.17E-03	0.10

Note: For each full model, we only report results from the algorithm with the highest AUC out of LR, SVM, XGBoost, and RF. We present the LR coefficients of each variable (Coef), the correlation of the variable with the outcome (Y-corr), the *p*-value, the mean of the variable (Y1-mean) in the intubated patients, and the mean of the variable (Y0-mean) in the nonintubated patients.

CRP: C-reactive protein; Total Elective Surgery percentage: (Total number of Elective Surgeries/Total number of beds)×100.

and White patients, avoiding to also examine bias involving Hispanic or Latino, which is another racial group with sufficient number of patients for such analysis. Several reasons for this choice: (1) there is considerable ambiguity on how people self-identify as Hispanic or Latino[33]; (2) in our data set about 44% of the patients have a missing race variable and the majority of those also identified as Hispanic/Latino (about 80%); and (3) there are disparities even between Black and White Hispanic/Latino individuals.[34]

Food insecurity and need for transportation became the top predictive features in the hospitalization model, possibly because they serve as a marker for severe economic hardship. Food is the most basic need and is related to patients' lifestyles and state of health. The COVID pandemic further expanded food insecurity worldwide, making it harder for vulnerable households to address their needs. Similarly, patients with transportation needs rely more on the most affordable public transit, which increases their risk for exposure to

the SARS-CoV-2 virus, while people with private cars and those who work from home can avoid exposure. Further, delayed access to care leads to a possibly worse clinical condition when arriving in acute settings. The other SDOH variables, such as housing insecurity were not as predictive as "Food" and "Transportation," possibly because the homeless rate in Boston has dropped sharply in recent years; specifically, 97–98% of the homeless population has been sheltered according to the latest homeless census.[35]

The percentage of Black patients in the data set is 35.1%, yet their percentage in the admitted, ICU, patients, mechanically ventilated, and deceased ranges from 43.1% to 45.5%. Predictive models exploit biases in the underlying data.[36] The hospitalization model exhibits bias, being more likely to falsely predict that a Black patient will be hospitalized. This reinforces the consideration of race as a social construct; persons who identify as Black being adversely affected by structural racism, and associated with a host of

Table 4. Mortality prediction models with features extracted within 3 days after admission

	Adm-based		24 h-drop	
	AUC	F1-weighted	AUC	F1-weighted
Best full models before SFS	XGBoost		XGBoost	
	91.4% (0.8%)	91.9% (0.7%)	96.2% (0.7%)	94.6% (0.6%)
Best full models after SFS	XGBoost		XGBoost	
	91.1% (2.0%)	91.8% (1.3%)	94.7% (1.2%)	94.3% (1.0)
Parsimonious models	92.0% (2.9%)	92.4% (1.3%)	94.0% (0.6%)	92.7% (0.9%)
LR-L1-LSTM-Transformer				
BMC-Protocol	88.9% (4.1%)	91.2% (0.7%)	89.5% (2.2%)	91.5% (0.6%)
LR-L1				
BMC-Protocol XGBoost	89.3% (2.4%)	90.8% (0.7%)	92.3% (1.1%)	92.6% (1.4%)
NEWS2 score	66.8% (5.0%)	85.8% (0.9%)	72.4% (2.9%)	87.2% (1.1%)
qSOFA score	68.2% (3.0%)	86.6% (0.9%)	73.3% (5.1%)	87.1% (0.6%)
LSTM-Transformer score	84.3% (4.4%)	89.0% (0.9%)	87.1% (2.0%)	91.0% (1.5%)

Variable	Coef	Y1-mean	Y0-mean	p-value	Y-corr
Adm-based parsimonious model 13 features					
LSTM-Transformer vital score	1.55	1.88	−3.32	0.00E + 00	0.53
Intercept	−1.98	–	–	–	–
Age	0.48	69.96	55.96	1.11E-16	0.23
CRP−3τ	0.36	140.83	80.83	1.55E-08	0.26
PMH-CAD	0.34	0.19	0.04	2.74E-11	0.19
Red Cell Distribution Width−1τ	0.31	15.56	14.27	3.06E-06	0.19
Poly Neutrophils−3τ	0.29	74.55	65.77	1.42E-09	0.22
Alkaline Phosphatase (total)−2τ	0.27	111.91	87.40	3.36E-03	0.13
Blood Component Type (Red Cell Group)−1τ	0.23	0.31	0.10	9.77E-12	0.20
Calcium−3τ	−0.22	8.51	8.76	4.85E-05	−0.16
Potassium−1τ	0.22	4.44	4.01	2.89E-15	0.23
Ferritin−1τ	0.16	2846.19	1026.75	2.41E-06	0.20
Platelets−2τ	−0.15	226.14	245.33	1.13E-02	−0.06
Glucose−2τ	0.13	149.06	129.09	4.39E-05	0.11
24 h-drop parsimonious model 13 features					
LSTM-Transformer vital score	2.67	2.81	−3.01	0.00E + 00	0.63
Intercept	−3.67	–	–	–	–
PH Arteria−1τ	−1.32	7.34	7.38	1.16E-12	−0.28
White Blood Cells−3τ	1.13	10.53	8.27	8.88E-16	0.21
Platelets−3τ	−1.07	230.18	253.84	4.71E-14	−0.11
LDH−4τ	0.93	691.01	537.40	2.57E-06	0.18
CRP−3τ	0.83	159.66	112.24	9.11E-10	0.27
Age	0.81	69.96	55.96	1.11E-16	0.23
PHM-CHF	0.67	0.23	0.08	6.86E-07	0.15
Absolute Lymphocytes−3τ	−0.59	1.29	1.45	5.06E-05	−0.03
Fibrinogen−3τ	−0.56	512.74	521.48	5.66E-04	−0.03
PCO$_2$ Arterial−2τ	0.42	46.14	42.43	1.27E-09	0.26
Poly Neutrophils−6τ	0.20	74.76	72.14	1.36E-09	0.18
Hemoglobin Blood−7τ	−0.10	10.88	11.16	4.41E-07	−0.15

Note: For each full model, we only report results from the algorithm with the highest AUC out of LR, SVM, XGBoost, and RF. We present the LR coefficients of each variable (Coef), the correlation of the variable with the outcome (Y-corr), the *p*-value, the mean of the variable (Y1-mean) in the deceased, and the mean of the variable (Y0-mean) in the nondeceased.

PMH: past medical history; CHF: congestive heart failure; CAD: coronary artery disease; CRP: C-reactive protein; LDH: lactate dehydrogenase.

circumstances, conditions and comorbidities that increase hospitalization risk.[37] As discussed earlier, it is possible to correct for this bias by employing different decision thresholds for Black and White patients.

Hospital census, such as the percentage of COVID-19 and non-COVID-19 patients and elective surgeries performed, affect the prediction results. This implies that an oversaturated hospital does affect resource allocation for new patients and further exacerbates the risk of future decompensation without adequate medical support.

Our hospitalization prediction model can be useful in any outpatient or emergency care setting given that the variables used are readily available to clinicians. This includes information on SDOH which is regularly collected at BMC. We note that such SDOH information gathering practices are becoming more widespread. Patients from underrepresented groups with potential SDOH needs are in fact more likely to present to an emergency care setting with ambulatory sensitive conditions compared to others.[38]

An FNR on the order of 0.25 achieved by the modified models corresponds to a reasonable compromise between false positive and

false negative decisions. Not hospitalizing/not transferring soon enough can lead to increased overall resource utilization as patients will present with more severe disease. This will have implications on the spread of the disease while they remain outside the hospital, length of the hospitalization and recovery if they end up being hospitalized, and associated economic consequences such as loss of wages. On the other hand, hospitalizing patients who may not need it leads to increased resource utilization and can result in hospitals been full and not having the ability to treat other patients who require care. Due to the novelty of COVID-19, including emerging variants, the related costs are not well characterized and vary greatly in different regions/hospitals, depending also on the local epidemiological conditions; therefore, hospitals may set this threshold based on their specific local situations.

A potential limitation of the study is that even though the hospitalization model has been externally validated, it has not been possible to do the same with the remaining models, particularly using data from other safety-net hospitals. In addition, although the patient past medical history data we used has no time limitations, underlying comorbidities may not be recorded in the EHR, which can potentially influence the performance or introduce bias in our models.

CONCLUSION

Our COVID-19 prediction models that are based on a large diverse patient population can accurately predict outcomes, potentially aiding in triage, resource allocation, and staffing determinations. Additionally, the use of dynamic variables such as vital signs improves the predictive ability of models and should be considered in future model development. This study highlights the importance of ensuring diverse patient populations are represented in advanced analytics development and suggests how to careful consider and interpret race within predictive models.

FUNDING

AUTHOR CONTRIBUTIONS

BH, YH, SS, and ZZ developed the models, obtained results, and co-wrote the manuscript. WGA, SAA, HS, and RGM provided access to data, medical intuition, contributed to the writing manuscript, and reviewed the manuscript. ICP designed/led the study, contributed to model development, and co-wrote the manuscript.

SUPPLEMENTARY MATERIAL

Supplementary material is available at *Journal of the American Medical Informatics Association* online.

CONFLICT OF INTEREST STATEMENT

None declared.

DATA AVAILABILITY

A data use agreement with the Boston Medical Center does not allow us to make the original data available. Code for the various algorithms that produced the results are available at https://github.com/noc-lab/BMC_COVID.

REFERENCES

1. Holder J. Tracking coronavirus vaccinations around the world. *The New York Times*, January 29, 2021. https://www.nytimes.com/interactive/2021/world/covid-vaccinations-tracker.html Accessed December 19, 2021.
2. Murthy BP. Disparities in COVID-19 vaccination coverage between urban and rural counties – United States, December 14, 2020–April 10, 2021. *MMWR Morb Mortal Wkly Rep* 2021; 70: 759–64. doi:10.15585/mmwr.mm7020e3
3. Peiffer-Smadja N, Lucet JC, Bendjelloul G, *et al*. Challenges and issues about organizing a hospital to respond to the COVID-19 outbreak: experience from a French reference centre. *Clin Microbiol Infect* 2020; 26 (6): 669–72.
4. Hilton CB, Milinovich A, Felix C, *et al*. Personalized predictions of patient outcomes during and after hospitalization using artificial intelligence. *NPJ Digit Med* 2020; 3 (1): 1–8.
5. Liang W, Liang H, Ou L, *et al*.; China Medical Treatment Expert Group for COVID-19. Development and validation of a clinical risk score to predict the occurrence of critical illness in hospitalized patients with COVID-19. *JAMA Intern Med* 2020; 180 (8): 1081–9.
6. Hao B, Sotudian S, Wang T, *et al*. Early prediction of level-of-care requirements in patients with COVID-19. *eLife* 2020; 9: e60519.
7. Yan L, Zhang HT, Goncalves J, *et al*. An interpretable mortality prediction model for COVID-19 patients. *Nat Mach Intell* 2020; 2 (5): 283–6.
8. Ji D, Zhang D, Xu J, *et al*. Prediction for progression risk in patients with COVID-19 pneumonia: the CALL Score. *Clin Infect Dis* 2020; 71 (6): 1393–9.
9. Wang K, Zuo P, Liu Y, *et al*. Clinical and laboratory predictors of in-hospital mortality in patients with coronavirus disease-2019: a cohort study in Wuhan, China. *Clin Infect Dis* 2020; 71 (16): 2079–88.
10. Gong J, Ou J, Qiu X, *et al*. A tool for early prediction of severe coronavirus disease 2019 (COVID-19): a multicenter study using the risk Nomogram in Wuhan and Guangdong, China. *Clin Infect Dis* 2020; 71 (15): 833–40. doi:10.1093/cid/ciaa443
11. Bhargava A, Fukushima EA, Levine M, *et al*. Predictors for severe COVID-19 infection. *Clin Infect Dis* 2020; 71 (8): 1962–8.
12. Bassett MT, Chen JT, Krieger N. Variation in racial/ethnic disparities in COVID-19 mortality by age in the United States: a cross-sectional study. *PLoS Med* 2020; 17 (10): e1003402.
13. Wollenstein-Betech S, Silva AAB, Fleck JL, Cassandras CG, Paschalidis IC. Physiological and socioeconomic characteristics predict COVID-19 mortality and resource utilization in Brazil. *PLoS One* 2020; 15 (10): e0240346.
14. Chowkwanyun M, Reed AL. Racial health disparities and Covid-19—caution and context. *N Engl J Med* 2020; 383 (3): 201–3.
15. Wynants L, Van Calster B, Collins GS, *et al*. Prediction models for diagnosis and prognosis of covid-19: systematic review and critical appraisal. *BMJ* 2020; 369: m1328.
16. Hochreiter S, Schmidhuber J. Long short-term memory. *Neural Comput* 1997; 9 (8): 1735–80.
17. Vaswani A, Shazeer N, Parmar N, *et al*. Attention is all you need. *Adv Neural Inf Process Syst* 2017; 30: 1–11.
18. BMC Facts. Boston, MA: Boston Medical Center; 2021. https://www.bmc.org/sites/default/files/For_Medical_Professionals/BMC-Facts.pdf Accessed December 23, 2021.
19. de la Vega PB, Losi S, Martinez LS, *et al*. Implementing an EHR-based screening and referral system to address social determinants of health in primary care. *Med Care* 2019; 57 (Suppl 2): S133–9.

20. Shi W, Tong L, Zhu Y, Wang MD. COVID-19 automatic diagnosis with radiographic imaging: explainable attention transfer deep neural networks. *IEEE J Biomed Health Inform* 2021; 25 (7): 2376–87.

21. Cortes C, Vapnik V. Support-vector networks. *Mach Learn* 1995; 20 (3): 273–97.

22. Chen T, Guestrin C. XGBoost: a scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. KDD '16. New York: Association for Computing Machinery (ACM); 2016: 785–794; San Francisco, CA.

23. Breiman L. Random forests. *Mach Learn* 2001; 45 (1): 5–32.

24. Chen R, Paschalidis IC. Distributionally robust learning. *Found Trends Optim* 2020; 4 (1–2): 1–243.

25. Chen R, Paschalidis IC, Hatabu H, Valtchinov VI, Siegelman J. Detection of unwarranted CT radiation exposure from patient and imaging protocol meta-data using regularized regression. *Eur J Radiol Open* 2019; 6: 206–11.

26. Smith G, Collins G, Gerry S, *et al.* The national early warning score 2 (NEWS2). *Clin Med* 2019; 19 (3): 260.

27. Seymour CW, Liu VX, Iwashyna TJ, *et al.* Assessment of clinical criteria for sepsis: for the Third International Consensus Definitions for Sepsis and Septic Shock (Sepsis-3). *Jama* 2016; 315 (8): 762–74.

28. Caton S, Haas C. Fairness in machine learning: a survey. Published online October 4, 2020. *arXiv:2010.04053*. doi: 10.48550/arXiv.2010.04053.

29. Rodriguez VA, Bhave S, Chen R, *et al.* Development and validation of prediction models for mechanical ventilation, renal replacement therapy, and readmission in COVID-19 patients. *J Am Med Inform Assoc* 2021; 28 (7): 1480–8. doi:10.1093/jamia/ocab029

30. Kantroo V, Kanwar MS, Goyal P, *et al.* Mortality and clinical outcomes among patients with COVID-19 and diabetes. *Med Sci* 2021; 9 (4): 65.

31. Huang I, Pranata R, Lim MA, Oehadian A, Alisjahbana B. C-reactive protein, procalcitonin, D-dimer, and ferritin in severe coronavirus disease-2019: a meta-analysis. *Ther Adv Respir Dis* 2020; 14: 17534666 20937175.

32. Henry BM, Aggarwal G, Wong J, *et al.* Lactate dehydrogenase levels predict coronavirus disease 2019 (COVID-19) severity and mortality: a pooled analysis. *Am J Emerg Med* 2020; 38 (9): 1722–6.

33. Allen VC Jr, Lachance C, Rios-Ellis B, Kaphingst KA. Issues in the assessment of "race" among Latinos: implications for research and policy. *Hisp J Behav Sci* 2011; 33 (4): 411–24.

34. LaVeist-Ramos TA, Galarraga J, Thorpe RJ, Bell CN, Austin CJ. Are black Hispanics black or Hispanic? Exploring disparities at the intersection of race and ethnicity. *J Epidemiol Community Health* 2012; 66 (7): e21.

35. 2020 Census Memo.pdf. Boston, MA. https://www.boston.gov/sites/default/files/file/2020/11/2020%20Census%20Memo.pdf Accessed February 24, 2022.

36. Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. *Science* 2019; 366 (6464): 447–53.

37. Khazanchi R, Evans CT, Marcelin JR. Racism, not race, drives inequity across the COVID-19 continuum. *JAMA Netw Open* 2020; 3 (9): e2019933.

38. Brown LE, Burton R, Hixon B, *et al.* Factors influencing emergency department preference for access to healthcare. *West J Emerg Med* 2011; 13 (5): 410–5.