



Research and Applications

Development and validation of prediction models for mechanical ventilation, renal replacement therapy, and readmission in COVID-19 patients

Victor Alfonso Rodriguez,^{1,*} Shreyas Bhavne,^{1,*} Ruijun Chen , Chao Pang,^{1,2} George Hripcsak,¹ Soumitra Sengupta,¹ Noemie Elhadad,¹ Robert Green,³ Jason Adelman,⁴ Katherine Schlosser Metitiri,⁵ Pierre Elias,¹ Holden Groves,⁶ Sumit Mohan,⁷ Karthik Natarajan  and Adler Perotte¹

¹Department of Biomedical Informatics, Columbia University, New York, New York, USA, ²Department of Translational Data Science and Informatics, Geisinger, Danville, Pennsylvania, USA, ³Department of Emergency Medicine, Columbia University Irving Medical Center, New York, New York, USA, ⁴Division of General Medicine, Department of Medicine, Columbia University Irving Medical Center, New York, New York, USA, ⁵Department of Pediatrics, Columbia University Irving Medical Center, New York, New York, USA, ⁶Department of Anesthesiology, Columbia University Irving Medical Center, New York, New York, USA, and ⁷Division of Nephrology, Department of Medicine, Columbia University Irving Medical Center, New York, New York, USA

*Co-first authors.

Corresponding Author: Victor Alfonso Rodriguez, MPhil, 622 West 168th St., PH20; New York, NY 10032, USA; victor.a.rodriguez@columbia.edu

Received 20 September 2020; Revised 9 January 2021; Editorial Decision 3 February 2021; Accepted 5 February 2021

ABSTRACT

Objective: Coronavirus disease 2019 (COVID-19) patients are at risk for resource-intensive outcomes including mechanical ventilation (MV), renal replacement therapy (RRT), and readmission. Accurate outcome prognostication could facilitate hospital resource allocation. We develop and validate predictive models for each outcome using retrospective electronic health record data for COVID-19 patients treated between March 2 and May 6, 2020.

Materials and Methods: For each outcome, we trained 3 classes of prediction models using clinical data for a cohort of SARS-CoV-2 (severe acute respiratory syndrome coronavirus 2)-positive patients ($n = 2256$). Cross-validation was used to select the best-performing models per the areas under the receiver-operating characteristic and precision-recall curves. Models were validated using a held-out cohort ($n = 855$). We measured each model's calibration and evaluated feature importances to interpret model output.

Results: The predictive performance for our selected models on the held-out cohort was as follows: area under the receiver-operating characteristic curve—MV 0.743 (95% CI, 0.682-0.812), RRT 0.847 (95% CI, 0.772-0.936), readmission 0.871 (95% CI, 0.830-0.917); area under the precision-recall curve—MV 0.137 (95% CI, 0.047-0.175), RRT 0.325 (95% CI, 0.117-0.497), readmission 0.504 (95% CI, 0.388-0.604). Predictions were well calibrated, and the most important features within each model were consistent with clinical intuition.

Discussion: Our models produce performant, well-calibrated, and interpretable predictions for COVID-19 patients at risk for the target outcomes. They demonstrate the potential to accurately estimate outcome prognosis in resource-constrained care sites managing COVID-19 patients.

Conclusions: We develop and validate prognostic models targeting MV, RRT, and readmission for hospitalized COVID-19 patients which produce accurate, interpretable predictions. Additional external validation studies are needed to further verify the generalizability of our results.

Key words: COVID-19, supervised machine learning, renal replacement therapy, respiration, artificial, patient readmission

INTRODUCTION

The United States continues to be a major epicenter for coronavirus disease 2019 (COVID-19), the disease caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2).^{1,2} In the early phase of the pandemic, hospitals in hard-hit regions, such as the New York Metropolitan Area, suffered large caseloads which heavily strained medical resources.³⁻⁵ The surge in cases throughout the country continues to drive medical resource expenditure, exhausting limited supplies. In this setting, delivering optimal care to COVID-19 patients will require matching scarce resources to patients in need across hospital systems, cities, and even across the country. Efficient distribution of resources will depend critically on prognostic assessments for newly presenting patients. With accurate prognostication, patient needs may be anticipated and met with the necessary equipment and provider expertise to limit disease progression or guard against avoidable adverse outcomes. In this study, we develop, validate, and analyze predictive models for the prognostication of 3 prevalent and actionable adverse outcomes in the setting of COVID-19.

Acute respiratory failure (ARF) requiring mechanical ventilation, severe acute kidney injury (AKI) requiring renal replacement therapy (RRT), and readmission are 3 common and critical adverse outcomes for patients with COVID-19. Roughly 12% to 33% of patients suffer ARF and require mechanical ventilation.⁶⁻⁹ 34% of all patients with COVID-19 and 78% of COVID-19 intensive care unit (ICU) patients develop AKI, with up to 14% of all patients and 35% of ICU patients requiring RRT.¹⁰ In addition, while hospitals struggle to manage heavy COVID-19 caseloads, patients who would normally be admitted may be discharged home, leading to higher than expected readmission rates.¹¹ Each of these outcomes carries significant implications for patient outcomes, long-term sequelae, and utilization of scarce resources including hospital beds and the equipment and materials needed for mechanical ventilation and RRT. Clinical prediction models could be used effectively to assess patient prognosis, informing resource planning and triage decisions.^{12,13} Nevertheless, most published COVID-19 prediction models have focused on disease diagnosis, while the few prognostic models have targeted COVID-19 disease severity or mortality.¹⁴

In this work, we aim to build interpretable prognostic models for COVID-19 patients that estimate the risk of ARF requiring mechanical ventilation, AKI requiring RRT, and hospital readmission. We develop our models using electronic health record (EHR) data from a major tertiary care center in New York City during the peak of the COVID-19 crisis, and externally validate them using data from a community hospital.

MATERIALS AND METHODS

Data sources and patient population

We focus on patients whose hospital courses included emergency room visits, inpatient admissions, or both at Columbia University Irving Medical Center/NewYork-Presbyterian (CUIMC/NYP) between March 2 and May 6, 2020. As we are interested in studying

patients with active SARS-CoV-2 infection, we further limit this cohort to patients with a positive, polymerase chain reaction–based SARS-CoV-2 test at any point during their hospital course. All clinical observations were extracted from CUIMC/NYP's Clinical Data Warehouse formatted according to the Observational Medical Outcomes Partnership (OMOP) common data model.¹⁵ Data from CUIMC including Milstein Hospital and the Morgan Stanley Children's Hospital were used for model development. Observations for patients treated at NYP Allen Hospital, a community hospital member of NYP, were held out as a validation set. We use chi-square permutation tests to compare the distribution of outcomes and demographics between our 2 cohorts (see [Table 1](#)).

We note that the development and validation cohort data are derived from care sites with distinct inpatient and critical care capacities. To characterize these differences, we provide each site's regular inpatient and ICU bed counts as well as their average annual admissions (see [Supplementary Table 1](#)).

Clinical observations

Our datasets comprise demographics, smoking status, laboratory test results, vital signs, and conditions. Clinical laboratory tests and vital signs are standardized while demographics and conditions are transformed into a binary encoding indicating presence (see [Supplementary Methods](#)). We include in our feature set only those conditions that appeared in the clinical records of at least 5 patients. A full list of the variables included in our feature set is provided in [Supplementary Table 2](#).

For RRT and mechanical ventilation models, we use data gathered during the first 12 hours of the current hospital course. The 12-hour constraint is meant to exclude early events, which are likely to be anticipated on presentation and are therefore less likely to be intervened upon based on the output from a predictive model. This constraint also removes episodes occurring prior to a patient's arrival at the hospital; such events must be excluded to permit construction of prognostic models. We also include data from patients' prior visits. For numerical data types like laboratory tests and vital signs, we use only the most recent values. Our binary encoding accounts for the presence of conditions in a patient's current visit and all their prior visits.

The dataset for our readmission models is constructed in the same way as is done for the mechanical ventilation and RRT models with one important difference: we extend the data-gathering period to cover the entirety of the index hospital admission, not just the first 12 hours.

Handling missing values

For conditions, the binary encoding does not require imputation as it encodes presence directly. For numerical variables, we impute missing values using Scikit-learn's¹⁶ implementation of the MICE algorithm¹⁷ with its default parameterization. Categorical variables (excluding conditions) were imputed using the most common class in the training set. Furthermore, for imputed variables, we expand our features to include binary missingness indicators specifying

Table 1. Characteristics and target outcomes for patients with SARS-CoV-2–positive tests

	Development (CUIMC) (n = 2256)	Validation (Allen Hospital) (n = 855)	P Value
Outcome			
Mechanical ventilation	352 (15.60)	60 (7.02)	<.0001
Renal replacement therapy	142 (6.29)	20 (2.34)	<.0001
Readmission	193 (8.55)	77 (9.01)	.7216
Age			<.0001
<18 y	50 (2.22)	0 (0)	
18-30 y	113 (5.01)	25 (2.92)	
30-60 y	761 (33.73)	242 (28.30)	
60-80 y	916 (40.60)	378 (44.21)	
>80 y	416 (18.44)	210 (24.56)	
Sex			.8987
Female	1005 (44.55)	375 (43.86)	
Male	1250 (55.41)	479 (56.02)	
Missing	1 (0.04)	1 (0.12)	
Race			.1275
American Indian or Alaska Native	3 (0.13)	1 (0.12)	
Asian	29 (1.29)	5 (0.58)	
Black or African American	455 (20.17)	192 (22.46)	
Native Hawaiian or Other Pacific Islander	10 (0.44)	0 (0)	
White	542 (24.02)	196 (22.92)	
Missing	1217 (53.95)	461 (53.92)	
Ethnicity			.0003
Hispanic or Latino	1068 (47.34)	439 (51.35)	
Not Hispanic or Latino	605 (26.82)	254 (29.71)	
Missing	583 (25.84)	162 (18.95)	
DNR/DNI	331 (14.67)	169 (19.76)	<.0001
Died in hospital	228 (10.11)	150 (17.54)	<.0001

Values are n (%).

CUIMC: Columbia University Irving Medical Center; DNI: do not intubate; DNR: do not resuscitate; SARS-CoV-2: severe acute respiratory syndrome coronavirus 2.

whether a value was observed or imputed. See [Supplementary Table 3](#) for a detailed account of each variable's missingness proportion.

Outcome definitions

We construct definitions for mechanical ventilation, RRT, and readmission, and constrain our analysis to the earliest such event within a patient's available timeline. Note that we do not exclude patients who died during their index hospital course. As such, our outcome-positive cohorts contain patients who experienced the target outcome and died afterward. Conversely, outcome-negative cohorts contain patients who died without ever experiencing the target outcome. This choice is in line with our aim of constructing clinically useful prognostic models. Doing so requires that we construct our models using data for all available patients, including those who deteriorate so quickly that they expire before our target outcomes can take place as well as those who deteriorate after all potential clinical interventions have been exhausted. See [Table 1](#) for summary mortality statistics.

We validate all our outcome definitions by iteratively sampling 50 to 100 patients, reviewing their clinical records to determine if our outcome definitions correctly classified their outcome status, and refining the outcome definitions to reduce misclassifications. Furthermore, we train preliminary models and review the clinical records for false positive and false negative patients to further revise our outcome definitions where appropriate.

Mechanical ventilation

In our Clinical Data Warehouse, structured data in electronic nursing flowsheets contain the most accurate observations and time-stamps regarding a patient's mechanical ventilation status. From these flowsheets, we extract the mechanical ventilation onset times for each patient in our cohort. If a patient undergoes multiple mechanical ventilation episodes within a single hospital course, we use the earliest onset time to identify the first such episode.

Renal replacement therapy

We use nursing flowsheets to extract the onset time of RRT for each patient and restrict to the earliest such episode. In addition, we exclude patients with a likely history of RRT by eliminating patients whose records contained OMOP concepts related to end-stage renal disease or stage 5 chronic kidney disease (see [Supplementary Table 4](#)).

Readmission

Readmissions were defined as any emergency visit or inpatient admission occurring 1 to 7 days after a previous emergency room or inpatient discharge. To calculate the interval between an individual patient's visits, we simply take the difference between the first and second visit's end time and start time, respectively. Readmissions occurring within 1 day postdischarge were excluded as these events are difficult to distinguish from transfers within an ongoing hospital stay. If multiple readmissions are observed, we focus on the one with the earliest start date.

Table 2. Performance metrics for all models and outcomes

Outcome	Model	AUROC (Development)	AUPRC (Development)	AUROC (Validation)	AUPRC (Validation)
Mechanical ventilation	Logistic L1	0.869 (0.847-0.893)	0.569 (0.510-0.624)	0.741 (0.682-0.806)	0.127 (0.052-0.157)
	Logistic EN	0.878 (0.858-0.902) <i>a</i>	0.562 (0.501-0.616)	0.738 (0.675-0.805)	0.141 (0.046-0.183) <i>a</i>
	GBT ^b	0.869 (0.848-0.891)	0.613 (0.555-0.668) <i>a</i>	0.743 (0.682-0.812) <i>a</i>	0.137 (0.047-0.175)
Renal replacement therapy	Logistic L1 ^b	0.847 (0.815-0.882) <i>a</i>	0.381 (0.293-0.453) <i>a</i>	0.847 (0.772-0.936) <i>a</i>	0.325 (0.117-0.497) <i>a</i>
	Logistic EN	0.844 (0.812-0.881)	0.378 (0.295-0.451)	0.841 (0.759-0.931)	0.314 (0.113-0.476)
	GBT	0.837 (0.805-0.871)	0.325 (0.242-0.385)	0.829 (0.761-0.912)	0.196 (0.009-0.312)
Readmission	Logistic L1	0.818 (0.789-0.847)	0.293 (0.233-0.344)	0.868 (0.823-0.917)	0.505 (0.395-0.602) <i>a</i>
	Logistic EN ^b	0.830 (0.803-0.858)	0.307 (0.249-0.353)	0.871 (0.830-0.917) <i>a</i>	0.504 (0.388-0.604)
	GBT	0.838 (0.814-0.864) <i>a</i>	0.287 (0.233-0.323)	0.869 (0.830-0.910)	0.427 (0.321-0.509)

AUROC: area under the receiver operating characteristic curve; AUPRC: area under the precision-recall curve; GBT: gradient boosted trees; Logistic EN: elastic-net logistic regression; Logistic L1: L1-penalized logistic regression.

^aThe best performance for the given outcome according to the metric specified by the column heading.

^bSelected models are in bold for each outcome.

Statistical analysis

Models

We employ 3 types of models: L1-penalized logistic regression (logistic L1), elastic-net logistic regression (logistic EN), and gradient boosted trees (GBT). The former 2 are based on logistic regression, an effective model for clinical prediction tasks.^{18,19} GBTs are nonparametric models that have also shown strong clinical prediction performance.²⁰⁻²³ These models are relatively simple, interpretable, and straightforward to apply for prognostic modeling. These characteristics align well with the aims of the present study. Furthermore, each of these models has a built-in regularization mechanism.²⁴⁻²⁷ This is crucial in our setting in which the number of features is on the order of the number patients. We use Scikit-learn to implement each model type.¹⁶ Both logistic L1 and logistic EN have a hyperparameter, *alpha*, which controls the strength of regularization. In addition, logistic EN has a second, *mixing* hyperparameter which controls the relative weight of the L1 vs L2 penalties. We use the default hyperparameter settings for GBT.

Model selection (hyperparameter tuning)

Our model selection approach relies upon 2 performance metrics: the area under the receiver-operating characteristic curve (AUROC) and the area under the precision-recall curve (AUPRC). For each model and outcome, we conducted 5-fold cross validation on the development cohort data searching across different hyperparameters (alpha: [0.3, 0.5, 0.7]; mixing: [1×10^{-4} , 1×10^4] equally spaced range of 10 values). We select the model with the best average AUROC across all folds. For the selected model, we compute the mean AUROC and AUPRC across all folds.²⁸ We obtain 95% confidence intervals (CIs) for all statistics by pooling the predicted probabilities and true labels across all folds within a reverse percentile bootstrap. For the validation cohort, we use the selected model to obtain outcome predictions and subsequently compute the reverse percentile bootstrap.

Calibration

For the development cohort, we use the pooled predicted probabilities and the true labels to generate the calibration curves. For the validation cohort, we use the predicted probabilities and true labels for the full cohort.

Feature importance

Feature importances for all models were evaluated using SHAP,²⁹ a method for estimating instance-wise Shapley values, which represent fair estimates of the effect each feature has upon an outcome predic-

tion. SHAP allows for instance-wise visualization, which for a given feature can demonstrate the distribution of the effect size and direction across the cohort.

Institutional review board

This study was approved by CUIMC's institutional review board and issued institutional review board number AAAS9678.

RESULTS

Cohort description

Our final development and validation cohorts contained 2256 and 855 patients, respectively. The distributions of outcome and demographic variables for each cohort are presented in Table 1. The distributions of sex and race and the number of readmissions were not significantly different between cohorts (*P* values >.05). Significant differences were found in the distributions of age and ethnicity, the numbers of mechanical ventilation and RRT cases, and the numbers of patients with do not intubate or do not resuscitate status or who died during their hospitalization (*P* < .001).

Model performance in development cohort

Performance metrics for all models and outcomes on the development cohort are presented in Table 2. The models with best AUROC for mechanical ventilation, RRT, and readmission were logistic EN (0.878 [95% CI, 0.858-0.902]), logistic L1 (0.847 [95% CI, 0.815-0.882]), and GBT (0.838 [95% CI, 0.814-0.864]), respectively. The best performing models according to AUPRC were GBT (0.613 [95% CI, 0.555-0.668]), logistic L1 (0.381 [95% CI 0.293-0.453]), and logistic EN (0.307 [95% CI, 0.249-0.353]), respectively.

Logistic L1 achieved the highest AUROC and AUPRC for RRT prediction; we use logistic L1 for all remaining RRT prediction experiments. No single model yielded the best performance on both metrics for mechanical ventilation and readmission. We chose GBT for mechanical ventilation and logistic EN for readmission, as they had the highest AUPRC and nearly highest AUROC (mechanical ventilation, 0.869 [95% CI, 0.848-0.891]; readmission, 0.830 [95% CI, 0.803-0.858]). See Figures 1 and 2 for development cohort ROC and precision-recall curves, respectively.

Model performance in validation cohort

Table 2 displays the performance metrics for all models on the validation cohort. Relative to the development cohort, mechanical ven-

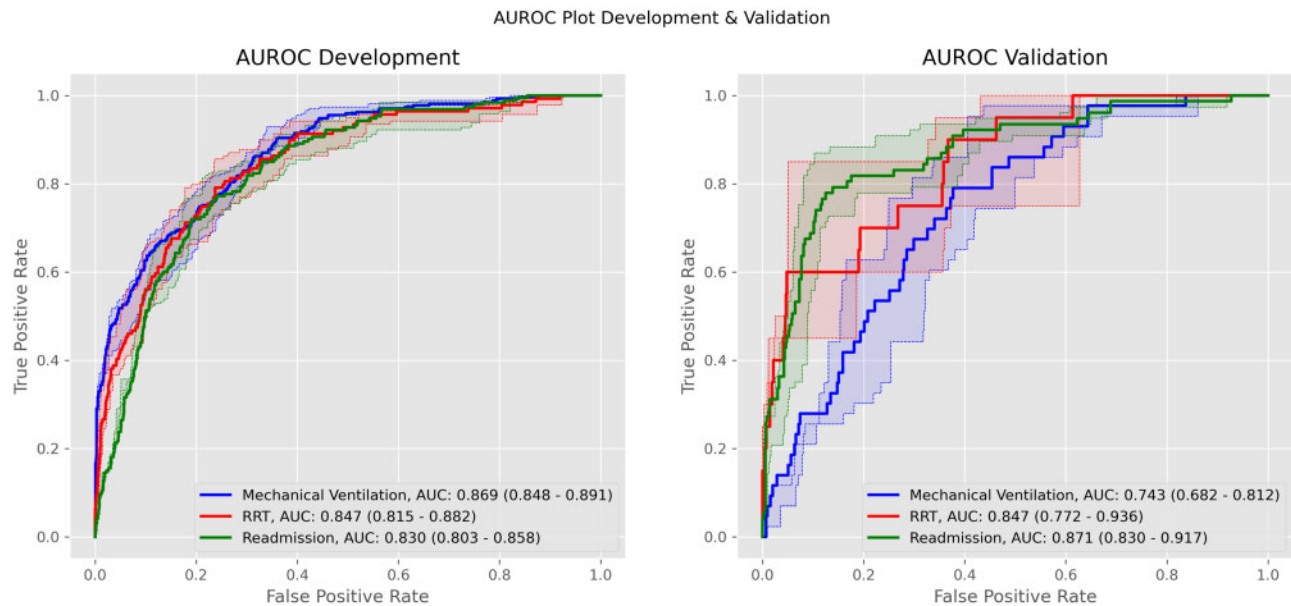


Figure 1. Receiver-operating characteristic (ROC) curves for ventilation, renal replacement therapy (RRT), and readmission. Curves are for each outcome's selected model. Dark lines correspond to averages over all folds. Shaded areas correspond to 95% confidence intervals. AUC: area under the curve; AUROC: area under the receiver-operating characteristic curve.

tilation predictive performance fell significantly across both AUROC (0.743 [95% CI, 0.682-0.812]) and AUPRC (0.137 [95% CI, 0.047-0.175]). RRT predictive performance remained consistent across both metrics (AUROC: 0.847 [95% CI, 0.772-0.936]; AUPRC: 0.325 [95% CI, 0.117-0.497]). For readmission, both metrics increased (AUROC: 0.871 [95% CI, 0.830-0.917]; AUPRC: 0.504 [95% CI, 0.388-0.604]). See [Figures 1](#) and [2](#) for validation cohort ROC and precision-recall curves, respectively.

Calibration

[Figure 3](#) shows the calibration curves for each outcome's selected model. In the development cohort, predicted probabilities for mechanical ventilation closely approximate the observed fraction of positive cases. Meanwhile, for both RRT and readmission, the predicted probabilities overestimate the fraction of positive cases. However, these estimates improve as the value of the predicted probability increases. Similar trends are observed for calibration in the validation cohort.

Feature importance

SHAP values for each outcome's selected model are visualized in [Figure 4](#). Respiratory illnesses including acute hypoxemic respiratory failure, acute respiratory distress syndrome (ARDS), and acute lower respiratory tract infection served as positive predictors (positive SHAP values) for mechanical ventilation. High respiratory rate, high neutrophil count, hypoxemia, shock, and documented disease due to coronaviridae (ie, the presence of the concept "Disease due to Coronaviridae" in a patient's clinical record) were also strong positive predictors. Greater age was negatively predictive (negative SHAP values).

Respiratory and renal illnesses including acute renal failure, acute hypoxemic respiratory failure, ARDS, and acute lower respiratory tract infection functioned as positive predictors for RRT. Several features drove the predicted likelihood either positively or negatively depending on their value. Serum creatinine, neutrophil count, C-reactive protein, and hyaline casts transition from nega-

tively to positively predictive as values increase from low to high. Meanwhile, serum bicarbonate and calcium make the same transition as values decrease. Furthermore, the presence of procalcitonin, urea nitrogen-to-creatinine ratio, and glomerular filtration rate measurements were positively predictive for RRT.

Readmission prediction was driven positively by high values for temperature, hemoglobin, and oxygen saturation (SpO₂). Conversely, it was driven negatively by low values for these variables. The opposite trend was observed for leukocyte count, respiratory rate, erythrocyte sedimentation rate (ESR), calcium, and erythrocyte distribution width. Fever and abdominal pain were positively predictive, whereas respiratory disorder and documented coronaviridae infection were negatively predictive. Missing values for laboratory tests including fibrin d-dimer, ferritin, procalcitonin, lactate dehydrogenase, ESR, and activated partial thromboplastin time were positively predictive.

DISCUSSION

Our results demonstrate that interpretable, performant, prognostic models targeting resource-intensive outcomes important to the management of COVID-19 may be trained using routinely recorded clinical variables. For mechanical ventilation and RRT, our models use only the data available within the first 12 hours of a patient's hospital course. Thus, their predictions may be made available to clinicians actively managing COVID-19 patients. Meanwhile, for readmission, our model utilizes data gathered throughout the current stay, making predictions available by the end of a hospital course when they would have the largest impact.

Our work extends and improves on the current state-of-the-art in outcome prediction for COVID-19 patients. Our mechanical ventilation prediction model is competitive with the deep learning model introduced by Shashikumar et al.³⁰ Though our objectives are distinct (their model targets hourly predictions), their validation AUROC (0.882) and AUPRC (0.209) lie near or within our 95% CIs. Our RRT prediction model demonstrates superior performance

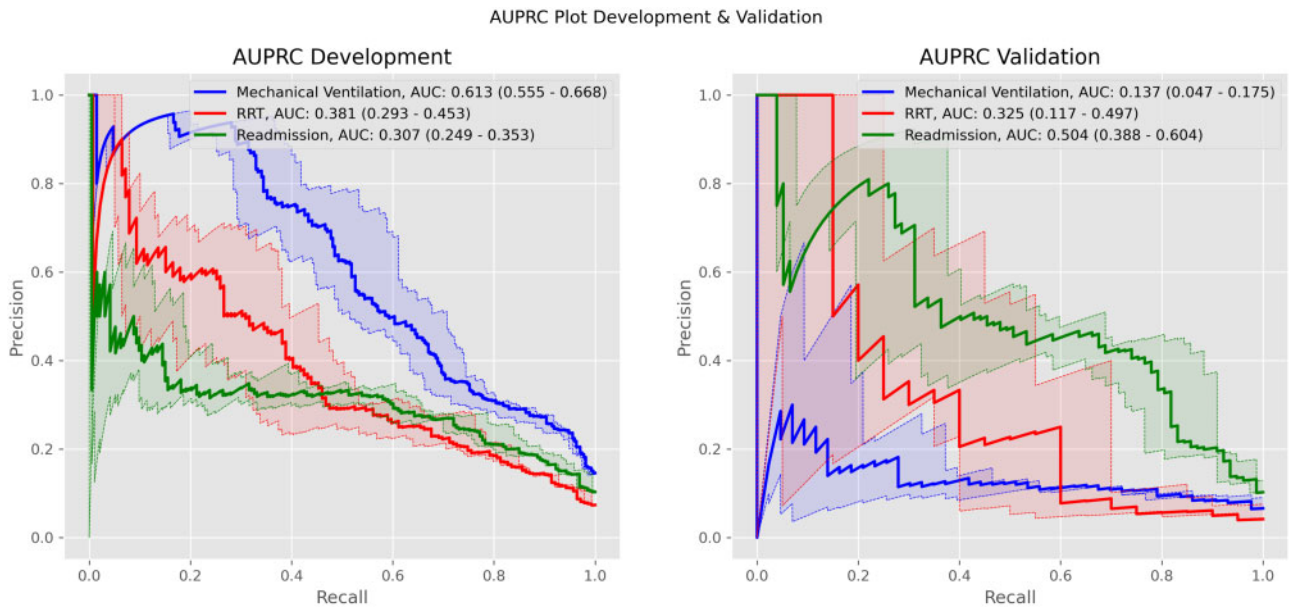


Figure 2. Precision-recall curves for ventilation, renal replacement therapy (RRT), and readmission. Curves are for each outcome’s selected model. Dark lines correspond to averages over all folds. Shaded areas correspond to 95% confidence intervals. AUC: area under the curve; AUPRC: area under the precision-recall curve.

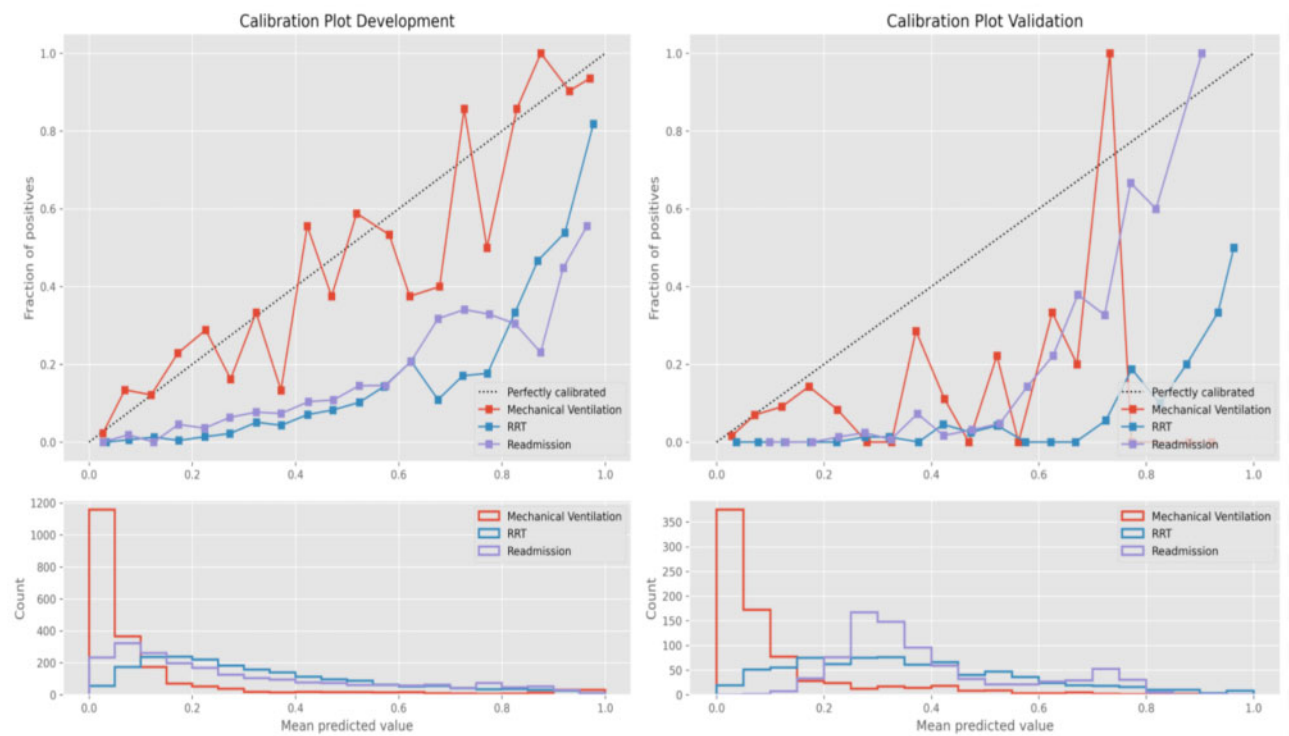


Figure 3. Calibration reliability curve for development and validation cohorts. The reliability curve shows how close each model is to a perfectly calibrated model. This plot is created by binning predicted probabilities and examining the true fraction of cases in each bin. The plot under each reliability curve shows the support (number of positives) in each bin. RRT: renal replacement therapy.

relative to previously described work which also utilized data from patients in New York City³¹; they obtained a validation AUROC of 0.79, which lies within our 95% CI (0.759-0.931). Though the current literature contains retrospective analysis studying the subpopulation of readmitted COVID-19 patients, to our knowledge, we are

the first to describe a predictive model for COVID-19 patient readmission.^{32,33}

Each of our models demonstrates reasonably good calibration in the development and validation cohorts. Nevertheless, caution should be taken when interpreting our models’ predicted probabili-

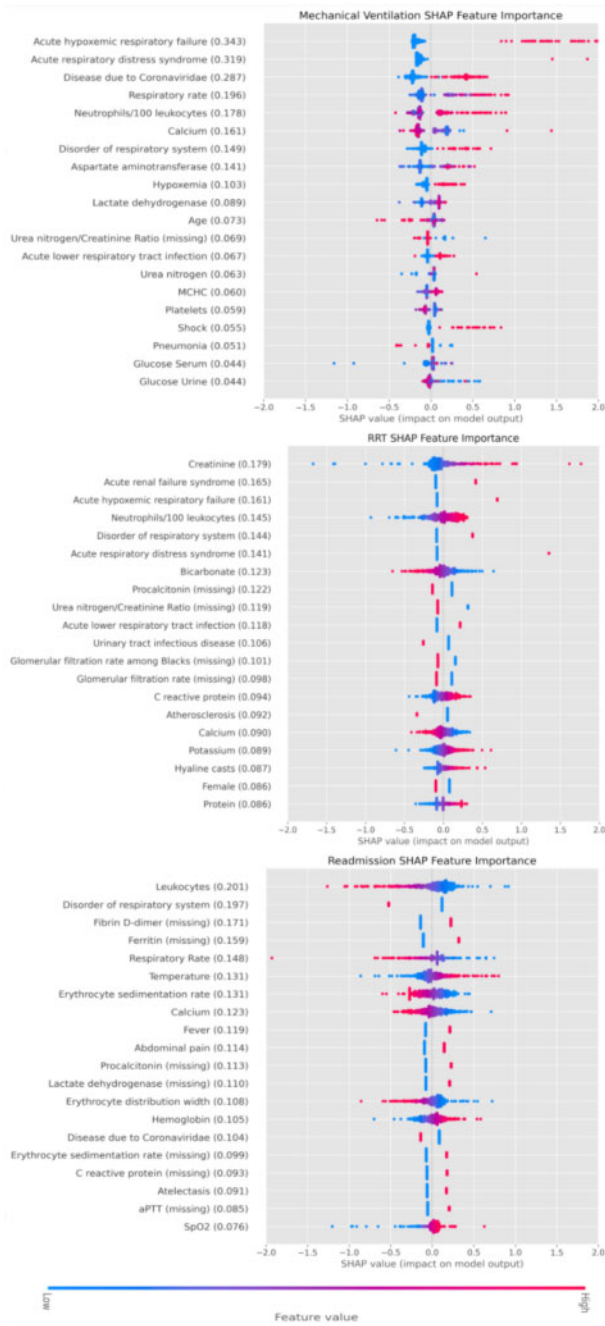


Figure 4. SHAP feature importances for ventilation, renal replacement therapy (RRT), and readmission. Each SHAP value plot displays a patient-level SHAP value as a point which lies on the horizontal axis and uses color to indicate whether the feature value for a patient was higher (red) or lower (blue) than average. SHAP values >1 indicate increased risk for a patient. SHAP values <1 indicate decreased risk. This SHAP plot allows for visualization of the distribution of effect sizes indicated by the spread of the points around 0 and shows the direction of the effect. As an example, a higher respiratory rate (red points are all >0) indicates higher risk for ventilation. The average of the absolute SHAP values (shown in parenthesis for each feature) across all points shows the overall importance of the feature. aPTT: activated partial thromboplastin time; MCHC: mean corpuscular hemoglobin concentration; SpO₂: oxygen saturation.

ties as estimates of the true risk of the target outcome for a given patient. Otherwise, a method for posttraining calibration should be employed, such as isotonic regression.³⁴

The use of SHAP values illuminates which features are driving our models' predictions and in which directions. This information is vital for evaluating what our models have learned. Consistent with expectations, predicted likelihoods of mechanical ventilation and RRT correlated positively with markers of respiratory and renal distress, as well as markers of active infectious or inflammatory processes. Notably, patient age was negatively predictive of mechanical ventilation, which is potentially a reflection of advance directives and clinical decision making, rather than a lower incidence of severe respiratory failure. In addition, as described in recently published work,³¹ we find that respiratory distress is strongly associated with RRT. Predicted probabilities for readmission were mostly driven by the absence of labs, which would be ordered if clinical suspicion for an infectious process were high (eg, lactate dehydrogenase, C-reactive protein, ESR). This finding suggests that readmitted patients may not have been considered ill enough to warrant admission and thus were given only a limited clinical workup for COVID-19. In addition, high respiratory rates were negatively predictive for readmission, suggesting that signs of respiratory distress may be associated with presentation later in the disease course, prolonged evaluation, and hence decreased probability of near-term return after discharge. Of note, the condition "Disease due to Coronaviridae" is a strong positive predictor for mechanical ventilation and a negative predictor for readmission. This suggests that the subset of patients whose documentation contains this concept code may be suffering from more severe disease on admission, as such patients are more likely to require invasive intervention (ie, mechanical ventilation) and are unlikely to be discharged early and be subsequently readmitted.

With further development and prospective validation, our outcome prediction models could potentially be utilized in practice to inform triage and resource allocation. Patients with high estimated risk of mechanical ventilation could be monitored more closely with continuous pulse oximetry and given early, noninvasive interventions such as self-proning.³⁵ Care could be taken to place these patients in beds with easy access to advanced oxygen therapies like high-flow nasal cannula and noninvasive positive pressure ventilation, resources that are typically not evenly distributed throughout a hospital. Similarly, providers could ensure that patients at high risk for RRT are placed in locations with the personnel and equipment needed to deliver this service. Such patients may also benefit from renal-protective therapeutic strategies such as setting a higher threshold for use of nephrotoxic agents, managing ARDS with less aggressive volume restriction, and an early nephrology consultation for AKI—an intervention that has been associated with improved renal prognosis.^{36,37} Additionally, given the relative paucity of dialysis equipment and appropriately trained staff during a pandemic surge, awareness of the risk of AKI requiring RRT could allow for improved resource planning and appropriate timing of surge protocols such as shared continuous RRT and acute peritoneal dialysis. Finally, patients with high risk of readmission could be re-evaluated for discharge, provided more intense monitoring, or provided additional support such as a visiting nurse that could help avoid a readmission while also lowering the risk of these patients decompensating at home.

Though our models demonstrate strong performance on the development cohort, we must also acknowledge that this performance deteriorates significantly when they are applied to the validation cohort. This observation speaks to the care practitioners must take when developing models on one patient population and applying them to another. In our case, the development cohort was drawn from a major medical center while the validation cohort came from

a small community hospital. It is likely that these 2 populations contain very different people who experience very different care practices. The result is development and validation datasets that differ in both the spectrum of observed variable values as well as the frequency and pattern of variable missingness. These differences limit our models' ability to generalize to the validation cohort what they learned on the development cohort and are likely a major driver of the performance degradation on the former. We also consider that differences in resource constraints, specifically regarding equipment and materials needed for mechanical ventilation and RRT, could have potentially contributed to our models' performance degradation on these outcomes in the validation cohort. However, due to changes in care practices (eg, having 2 patients share a single continuous RRT machine within a 24-hour on/off cycle) and acquisition of additional materials and equipment, neither site ever came close to exhausting its supplies. This suggests that resource constraints played at most a marginal and indirect role in limiting our models' performance on the validation cohort.

We acknowledge several important limitations to this work. Our data were derived from a single hospital network. This limits the generalizability of our results to other institutions, as we cannot capture the out-of-network variability in COVID-19 population characteristics and care practices. This limitation extends to our validation experiments, which used data from an in-network community hospital. This also complicates our modeling of readmission. Our positive readmission cases are limited to those patients whose discharge and readmission both occurred in our hospital network. Discharged patients who were subsequently admitted elsewhere would appear as negative cases in our models. We adopted a feature-agnostic approach when choosing which variables to include in our model. This allowed us to model many of the observations in the clinical record, but it also complicates the models' utility. To extract risk estimates from our model, a user will need to replicate our feature engineering and apply it to their local data stores. Thus, they will likely need a pipeline inputting clinical observations directly into the model from the EHR. Modeling many variables also introduced a significant amount of missing values (see [Supplementary Table 3](#)). To handle these, we used imputation strategies like MICE, which assume that the data are missing at random, even though our data are likely missing not at random. As such, it is likely that our fitted model parameters are biased.³⁸ However, as we are primarily concerned with optimizing prediction, we are willing to trade off model parameter bias for predictive performance by modeling the imputed data along with the observed missingness pattern.³⁹ Our use of the OMOP common data model also introduced challenges and limitations. The first of these concerns our outcome definitions, which relied on structured fields in nursing flowsheet. As these data are not part of the OMOP common data model, replicating our definitions at other sites may be difficult. Second, during the extraction, transformation, and loading of data into the OMOP we may have lost some observations. This is a likely source of the unusually large amount of missingness in routinely collected clinical measurements such as vitals and plasma and serum electrolyte labs.

CONCLUSION

In conclusion, we have trained and validated prognostic models targeting 3 significant, resource-intensive outcomes in the context of COVID-19: mechanical ventilation, RRT, and hospital readmission. Our models run on routinely collected clinical variables, and produce accurate, interpretable predicted likelihoods for each outcome.

Additional external validation studies are needed to further verify the generalizability of our results.

FUNDING

VAR is supported by grant F31LM012894 from the National Institutes of Health (NIH), National Library of Medicine (NLM). SB is supported by grant 5T15LM007079 from the NIH, NLM. AP is supported by grant R01HL148248 from the NIH, National Heart, Lung, and Blood Institute. The funders played no direct role in the present work with regard to study concept and design; acquisition, analysis, and interpretation of data; statistical analysis; manuscript drafting and revision; and supervision.

AUTHOR CONTRIBUTIONS

AP, KN, RC, SM, and HG contributed to concept and design. SB, CP, AP, VAR, KN, and SS contributed to acquisition, analysis, and interpretation of data. SB contributed to statistical analysis. VAR, SB, and RC contributed to manuscript drafting. VAR, RC, SM, NE, GH, KSM, JA, PE, RG, and HG contributed to manuscript revision. AP was involved in supervision.

SUPPLEMENTARY MATERIAL

[Supplementary material](#) is available at *Journal of the American Medical Informatics Association* online.

ACKNOWLEDGMENTS

The following authors had full access to the electronic health record data used in this work: AP, KN, CP, SB, VAR, and RC.

CONFLICTS OF INTEREST

The authors have no conflicts of interest to report as pertains to their financial interests, activities, relationships, and affiliations.

REFERENCES

- Centers for Disease Control and Prevention. Coronavirus disease 2019 (COVID-19). <https://www.cdc.gov/coronavirus/2019-ncov/cases-updates/cases-in-us.html> Accessed August 11, 2020.
- Johns Hopkins University and Medicine. Coronavirus Resource Center. <https://coronavirus.jhu.edu/>. Accessed August 11, 2020.
- Goldfarb DS, Benstein JA, Zhdanova O, *et al*. Impending shortages of kidney replacement therapy for covid-19 patients. *Clin J Am Soc Nephrol* 2020; 15 (6): 880–2.
- Ranney ML, Griffeth V, Jha AK. Critical supply shortages - the need for ventilators and personal protective equipment during the Covid-19 pandemic. *N Engl J Med* 2020; 382 (18): e41.
- Sanger-Katz M, Kliff S, Parlapiano A. These places could run out of hospital beds as coronavirus spreads. *The New York Times*. <https://www.nytimes.com/interactive/2020/03/17/upshot/hospital-bed-shortages-coronavirus.html> Accessed August 11, 2020.
- Argenziano MG, Bruce SL, Slater CL, *et al*. Characterization and clinical course of 1000 patients with coronavirus disease 2019 in New York: retrospective case series. *BMJ* 2020; 369: m1996.doi:10.1136/bmj.m1996.
- Goyal P, Choi JJ, Pinheiro LC, *et al*. Clinical characteristics of COVID-19 in New York City. *N Engl J Med* 2020; 382 (24): 2372–4.
- Richardson S, Hirsch JS, Narasimhan M, *et al*; Northwell COVID-19 Research Consortium. Presenting characteristics, comorbidities, and outcomes among 5700 patients hospitalized with COVID-19 in the New York City Area. *JAMA* 2020; 323 (20): 2052–9.
- Petrilli CM, Jones SA, Yang J, *et al*. Factors associated with hospital admission and critical illness among 5279 people with coronavirus disease

- 2019 in New York City: Prospective cohort study. *BMJ* 2020; 369: m1966.doi:10.1136/bmj.m1966.
10. Robbins-Juarez SY, Qian L, King KL, *et al.* Outcomes for patients with COVID-19 and acute kidney injury: a systematic review and meta-analysis. *Kidney Int Rep* 2020; 5 (8): 1149–60.
 11. Parra LM, Cantero M, Morras I, *et al.* Hospital readmissions of discharged patients with COVID-19. *Int J Gen Med* 2020; 13: 1359–66.
 12. Obermeyer Z, Emanuel EJ. Predicting the future-big data, machine learning, and clinical medicine. *N Engl J Med* 2016; 375 (13): 1216–9.
 13. Shilo S, Rossman H, Segal E. Axes of a revolution: challenges and promises of big data in healthcare. *Nat Med* 2020; 26 (1): 29–38.
 14. Cummings MJ, Baldwin MR, Abrams D, *et al.* Epidemiology, clinical course, and outcomes of critically ill adults with COVID-19 in New York City: a prospective cohort study. *Lancet* 2020; 395 (10239): 1763–70.
 15. Overhage JM, Ryan PB, Reich CG, Hartzema AG, Stang PE. Validation of a common data model for active safety surveillance research. *J Am Med Inform Assoc* 2012; 19 (1): 54–60.
 16. Pedregosa F, Varoquaux G, Gramfort A, *et al.* Scikit-learn: machine learning in Python. *J Mach Learn Res* 2011; 12: 2825–30.
 17. van Buuren S, Groothuis-Oudshoorn K. mice: Multivariate imputation by chained equations in R. *J Stat Softw* 2011; 45 (3): 1–68.
 18. Christodoulou E, Ma J, Collins GS, Steyerberg EW, Verbakel JY, Van Calster B. A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. *J Clin Epidemiol* 2019; 110: 12–22.
 19. Pavlou M, Ambler G, Seaman S, De Iorio M, Omar RZ. Review and evaluation of penalised regression methods for risk prediction in low-dimensional data with few events. *Stat Med* 2016; 35 (7): 1159–77.
 20. Atkinson EJ, Therneau TM, Melton LJ, *et al.* Assessing fracture risk using gradient boosting machine (GBM) models. *J Bone Miner Res* 2012; 27 (6): 1397–404.
 21. Ayaru L, Ypsilantis PP, Nanapragasam A, *et al.* Prediction of outcome in acute lower gastrointestinal bleeding using gradient boosting. *PLoS One* 2015; 10 (7): e0132485.
 22. Blagus R, Lusa L. Gradient boosting for high-dimensional prediction of rare events. *Comput Stat Data Anal* 2017; 113: 19–37.
 23. Sinha P, Churpek MM, Calfee CS. Machine learning classifier models can identify acute respiratory distress syndrome phenotypes using readily available clinical data. *Am J Respir Crit Care Med* 2020; 202 (7): 996–1004.
 24. Tibshirani R. Regression shrinkage and selection via the Lasso. *J R Stat Soc Series B Stat Methodol* 1996; 58 (1): 267–88.
 25. Zou H, Hastie T. Regularization and variable selection via the elastic net. *J R Stat Soc Series B Stat Methodol* 2005; 67 (2): 301–20.
 26. Hastie T, Tibshirani R, Wainwright M. *Statistical Learning with Sparsity: The Lasso and Generalizations*. Boca Raton, FL: CRC Press; 2015.
 27. Friedman JH. Greedy function approximation: a gradient boosting machine. *Am Stat* 2001; 29 (5): 1189–232.
 28. Ledell E, Petersen M, Van Der Laan M. Computationally efficient confidence intervals for cross-validated area under the ROC curve estimates. *Electron J Stat* 2015; 9 (1): 1583–607.
 29. Lundberg SM, Lee SI. A unified approach to interpreting model predictions. *Adv Neural Inform Process Syst* 2017; 4765–74.
 30. Shashikumar SP, Wardi G, Paul P, *et al.* Development and prospective validation of a transparent deep learning algorithm for predicting need for mechanical ventilation. *Chest* 2020 Dec 17 [E-pub ahead of print]. doi:10.1101/2020.05.30.20118109
 31. Chan L, Chaudhary K, Saha A, *et al.* AKI in hospitalized patients with COVID-19. *J Am Soc Nephrol*. 2021; 32 (1): 151–60.
 32. Somani S, Richter F, Fuster V, *et al.* Characterization of patients who return to hospital following discharge from hospitalization for COVID-19. *J Gen Intern Med* 2020; 35 (10): 2838–44
 33. Wang X, Xu H, Jiang H, *et al.* The clinical features and outcomes of discharged coronavirus disease 2019 patients a prospective cohort study. *QJM* 2020; 113 (9): 657–65.
 34. Chakravarti N. Isotonic median regression: a linear programming approach. *Math Oper Res* 1989; 14 (2): 303–8.
 35. Caputo ND, Strayer RJ, Levitan R. Early self-proning in awake, non-intubated patients in the emergency department: a single ED's experience during the COVID-19 pandemic. *Acad Emerg Med* 2020; 27 (5): 375–8.
 36. Ponce D, de Pietro Franco Zorzenon C, dos Santos NY, Balbi AL. Early nephrology consultation can have an impact on outcome of acute kidney injury patients. *Nephrol Dial Transplant* 2011; 26 (10): 3202–6.
 37. Soares DM, Pessanha JF, Sharma A, Brocca A, Ronco C. Delayed nephrology consultation and high mortality on acute kidney injury: a meta-analysis. *Blood Purif* 2017; 43 (1–3): 57–67.
 38. Rubin DB. Inference and missing data. *Biometrika* 1976; 63 (3): 581–92.
 39. Sperrin M, Martin GP, Sisk R, Peek N. Missing data should be handled differently for prediction than for description or causal explanation. *J Clin Epidemiol* 2020; 125: 183–7. doi:10.1016/j.jclinepi.2020.03.028