OPEN

# Development and Validation of ARC, a Model for Anticipating Acute Respiratory Failure in Coronavirus Disease 2019 Patients

**Suchi Saria, PhD**[1–3]

**Peter Schulam, PhD**[3]

**Brian J. Yeh, PhD**[3]

**Daniel Burke, MD, MBA**[3,4]

**Sean D. Mooney, PhD**[5]

**Christine T. Fong, MS**[6]

**Jacob E. Sunshine, MD**[6]

**Dustin R. Long, MD**[6]

**Vikas N. O'Reilly-Shah, MD, PhD**[6,7]

**OBJECTIVES:** To evaluate factors predictive of clinical progression among coronavirus disease 2019 patients following admission, and whether continuous, automated assessments of patient status may contribute to optimal monitoring and management.

**DESIGN:** Retrospective cohort for algorithm training, testing, and validation.

**SETTING:** Eight hospitals across two geographically distinct regions.

**PATIENTS:** Two-thousand fifteen hospitalized coronavirus disease 2019–positive patients.

**INTERVENTIONS:** None.

**MEASUREMENTS AND MAIN RESULTS:** Anticipating Respiratory failure in Coronavirus disease (ARC), a clinically interpretable, continuously monitoring prognostic model of acute respiratory failure in hospitalized coronavirus disease 2019 patients, was developed and validated. An analysis of the most important clinical predictors aligns with key risk factors identified by other investigators but contributes new insights regarding the time at which key factors first begin to exhibit aberrency and distinguishes features predictive of acute respiratory failure in coronavirus disease 2019 versus pneumonia caused by other types of infection. Departing from prior work, ARC was designed to update continuously over time as new observations (vitals and laboratory test results) are recorded in the electronic health record. Validation against data from two geographically distinct health systems showed that the proposed model achieved 75% specificity and 77% sensitivity and predicted acute respiratory failure at a median time of 32 hours prior to onset. Over 80% of true-positive alerts occurred in non-ICU settings.

**CONCLUSIONS:** Patients admitted to non-ICU environments with coronavirus disease 2019 are at ongoing risk of clinical progression to severe disease, yet it is challenging to anticipate which patients will develop acute respiratory failure. A continuously monitoring prognostic model has potential to facilitate anticipatory rather than reactive approaches to escalation of care (e.g., earlier initiation of treatments for severe disease or structured monitoring and therapeutic interventions for high-risk patients).

**KEY WORDS:** coronavirus disease 2019; deterioration monitoring; electronic surveillance; predictive model

Acute respiratory failure (ARF) remains a dangerous hallmark of hospitalized patients infected with severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), requiring careful monitoring and sometimes rapid intervention for optimal management (1). Identifying which patients will progress to ARF is a challenging and imprecise clinical task and has several substantive ramifications. Early escalations of care lead to inefficient use of system resources and unnecessary patient exposures to medications and

Copyright © 2021 The Authors. Published by Wolters Kluwer Health, Inc. on behalf of the Society of Critical Care Medicine. This is an open-access article distributed under the terms of the Creative Commons Attribution-Non Commercial-No Derivatives License 4.0 (CCBY-NC-ND), where it is permissible to download and share the work provided it is properly cited. The work cannot be changed in any way or used commercially without permission from the journal.

**DOI: 10.1097/CCE.0000000000000441**

invasive procedures (2). Conversely, failure to appreciate impending patient decline increases the risk of exposure for the clinical teams (e.g., due to the need for urgent donning of personal protective equipment) and risk of delayed care for patients (including late initiation of disease-modifying treatment, delayed goals of care discussion, and performance of procedures such as intubation under urgent circumstances). Under surge conditions, optimal patient disposition improves resource utilization and eases triaging of resources (e.g., nursing and respiratory therapy staff or monitoring equipment) in hospital systems under strain (3).

Anticipatory and prognostic models of progression to ARF in coronavirus disease 2019 (COVID-19) patients may help identify those at highest risk for deterioration, ameliorating aforementioned risks to the patient, care team, and hospital system. Many prognostic models of decompensation due to COVID-19 have been developed, although a recent review found many to be “poorly reported, at high risk of bias, and... probably optimistic” (4). Even so, models designed to assess risk at the time of admission have been found to aid with initial patient triage and key decisions regarding hospital admission or outpatient monitoring (5–8). However, a key gap in this literature is a lack of predictive models describing the ongoing risk of postadmission disease progression. The significant cohort of patients not admitted to the critical care setting at presentation represents a serious source of uncertainty about demand and outcomes for care teams and system-wide resources, uncertainty that predictive insight may diminish.

In the present work, we develop and validate prognostic COVID-19 models to fill this gap. Designed to model risk of progression to ARF following non-ICU inpatient admission, the models were trained using data from a health system in the Mid-Atlantic region of the United States and validated on data from an unrelated, geographically distinct health system in the Pacific Northwest (University of Washington [UW]). The main model, which we refer to as ARC (Anticipating Respiratory failure in Coronavirus disease), harnesses the full trajectory of vital signs and laboratory test results (referred to collectively as markers) following hospital admission. This model identifies the most important markers for predicting risk of ARF, defined by a requirement for substantial respiratory support (≥ 15 L/min, noninvasive positive pressure ventilation [NIPPV], or mechanical ventilation), and characterizes the timing between the moment these marker trajectories exhibit aberrency and the onset of acute respiratory deterioration. We also compared the performance of this model with one developed for ARF arising from non-ICU admission for pneumonia, findings with implications for the ongoing debate around specific differences in COVID-19 respiratory failure (9), as well as the optimal approaches to monitoring and clinical management of these patients when immediate ICU care is not required (10–12).

## MATERIALS AND METHODS

The focus of this study is the clinical progression of patients admitted to the hospital with confirmed SARS-CoV-2 who did not immediately require ICU-level care. Our objective is to identify clinical markers that are most predictive of impending ARF and to characterize the timing of observed abnormalities in the marker trajectories relative to the onset of ARF. To that end, we trained prognostic models of ARF in COVID-19 patients and characterized the risk models’ behavior longitudinally over the course of the encounter. We validated our findings on patients with confirmed COVID-19 from two health systems and compared marker contribution in COVID-19 with importance for ARF in other forms of pneumonia.

### Study Population

Our study included two independent cohorts of patients with confirmed COVID-19 (consistent with clinical practice, defined as a “positive” or “inconclusive” result from an reverse transcriptase-polymerase chain reaction test). Briefly, we excluded patients who experienced ARF in the emergency department or were admitted directly to the ICU. We also excluded patients treated in surgical care units and patients transferred to the ICU without ARF. After exclusions, the Mid-Atlantic cohort consisted of deidentified data from 1,741 patients with COVID-19 from five hospitals in a large medical center in the Mid-Atlantic region admitted between April 1, 2020, and July 22, 2020. The UW cohort consisted of deidentified data from 274 patients with confirmed COVID-19 from the three hospitals comprising the UW Medicine system (Seattle, WA) admitted between February 1, 2020, and July 9, 2020. A separate pneumonia cohort consisted of deidentified data from 3,475 patients from the five hospitals in the Mid-Atlantic health system that were hospitalized between February 1, 2014, and January 1, 2019. Additional details about the study populations and data extracted from each population are provided in the **Supplemental Methods** and **Tables S1–S3** (http://links.lww.com/CCX/A635). The study was approved by the UW Human Subjects Division Institutional Review Board (IRB)–D (IRB ID 00009977). This report adheres to the transparent reporting of a multivariable prediction model for individual prognosis or diagnosis and minimum information about clinical artifical intelligence modeling checklists for improved reporting of prediction model development and validation (13, 14). Data and source code will be made available upon reasonable request and execution of a Data Use Agreement (to ensure deidentified data are not reidentified).

## Outcome Definition

The World Health Organization's criteria for severe COVID-19 includes patients that are ever treated with high-flow supplemental oxygen, a noninvasive positive pressure device, or mechanical ventilator (15). Operationalizing the definition of high-flow required us to choose a cut off for numerical oxygen flow rate; as the maximal flow rate for conventional low-flow oxygen therapy is 15 L/min, we felt this was a reasonable breakpoint (16). In our cohorts, many patients were placed on 15 L/min of oxygen during brief sessions with non-physician providers (e.g., physical therapy). To mitigate false positives associated with the routine brief use of escalated therapy for potentially strenuous activities, we required at least one of the following criteria subsequent to the time when the patient is treated with 15 L/min or more of oxygen: 1) the patient is treated with 15 L/min or more of oxygen for more than 8 consecutive hours, 2) the patient is escalated to 30 L/min of oxygen or more, and 3) the patient is escalated to an NIPPV device or mechanical ventilator. If the patient was transiently treated with 15 L/min or more of oxygen and at least one of the above occurred, we considered the onset time of ARF to be when supplemental oxygen was raised to 15 L/min or more. If none of the additional criteria were met before oxygen was reduced to less than 15 L/min, but the additional criteria are met later in the same encounter, we considered the onset time to be the later time supplemental oxygen was raised to 15 L/min or more, or they required NIPPV or mechanical ventilation.

## Model Development

Detailed methods for model development and evaluation are provided in the Supplemental Methods (http://links.lww.com/CCX/A635). Briefly, we assigned hourly labels for each patient: for patients who did develop ARF, all hours within 24 hours prior to onset time were labeled as positive samples, and all prior hours were defined as negative samples. For patients who did not develop ARF, all hours were labeled as negative samples. For all markers, we calculated hourly predictors using the complete history of each marker until that time point. When a marker was not recorded prior to the timepoint, all features related to that marker were assigned a distinct nonnumeric value indicating missingness.

We fit all models in this article using gradient boosted decision trees as implemented in the Python xgboost package version 1.0.2 (17). We constructed training samples by subsampling each encounter. For each negative encounter, we randomly selected a single hour from the encounter and used the hourly predictors at that time as model input and the hourly label as the supervised output. For each positive encounter, we randomly selected a single positive hour (i.e., one within 24 hr of ARF onset), and if the patient was hospitalized for more than a day, a single negative hour (i.e., one outside of the 24 hr window prior to ARF). For each marker, we fit three different models. The full model used all features for a given marker, trained on the Mid-Atlantic data set. The latest value model used only the latest value for a given marker, trained on the Mid-Atlantic data set. The pneumonia model used all features for a given marker, trained on the pneumonia data set. We also trained three models using all features for all markers: a model trained on the pneumonia cohort, a model trained on the Mid-Atlantic cohort using markers available in the pneumonia cohort, and the final ARC model using all markers available in the Mid-Atlantic cohort. Tuning parameters and cross validation performance for the ARC model are shown in **Table S8** (http://links.lww.com/CCX/A636).

## Model Evaluation

We evaluated the discriminative power of each model using the area under the receiver operating characteristic curve (ROC AUC) at the encounter level. For patients who did not develop ARF, we calculated the

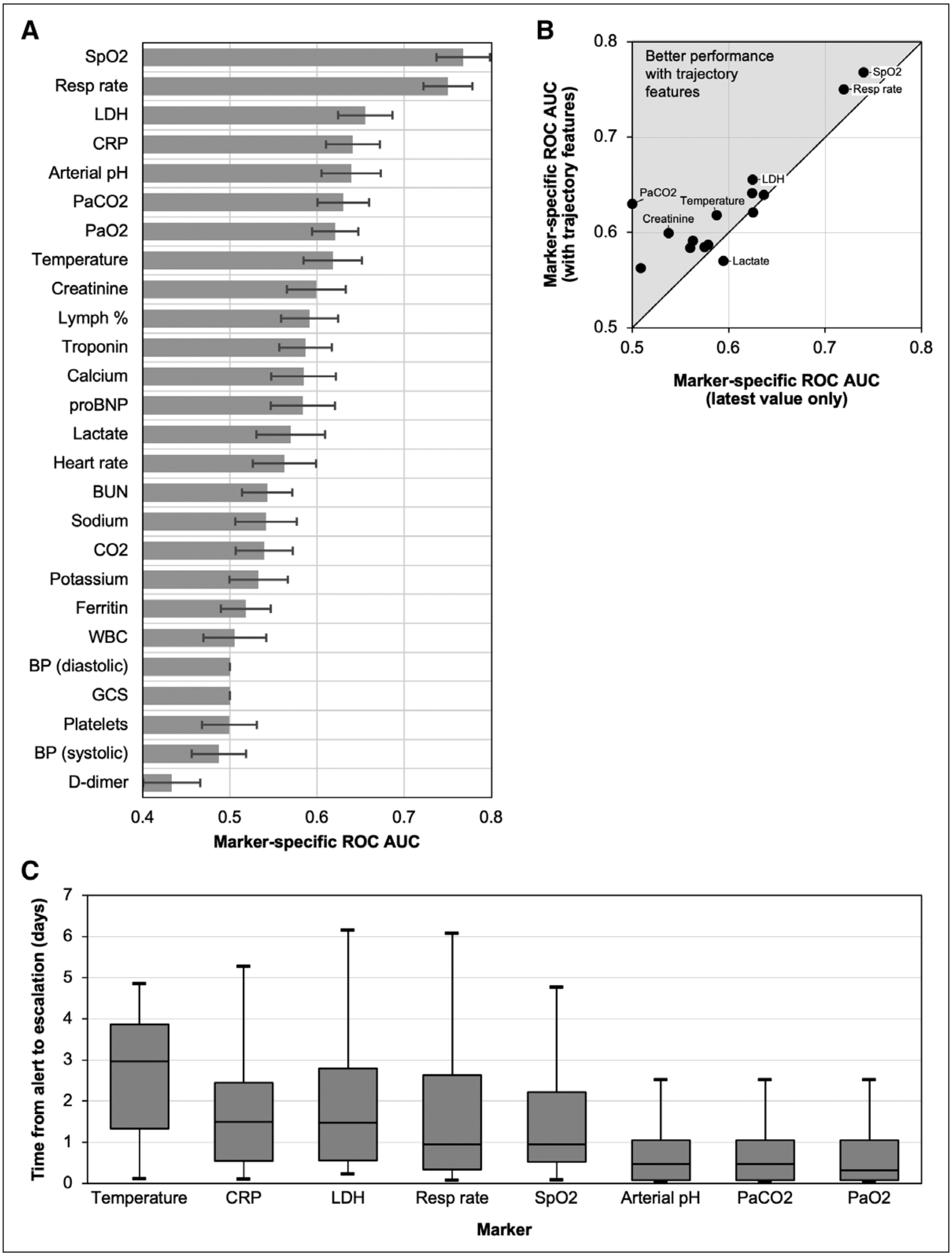

**Figure 1.** Prognostic value of markers and marker alert timing relative to acute respiratory failure (ARF). **A**, Receiver operating characteristic (ROC) area under the curve (AUC) for full models fit with individual markers and tested on pooled Mid-Atlantic test patients and University of Washington patients. AUC point estimates and SEs were computed using bootstrap resampling of encounters. **B**, Comparison of ROC AUC for marker-specific models incorporating trajectory features (*vertical axis*) against models incorporating only the latest value of the marker (*horizontal axis*). *Shaded area* represents markers where adding trajectory features improve performance. **C**, Timeliness of marker-specific models relative to onset of ARF. *Box plot* depicts the difference between alert onset (time when a given model score first crosses a threshold selected for 75% specificity) and onset of ARF. Markers are organized in order of decreasing median time between alert onset and ARF onset. BP = blood pressure, BUN = blood urea nitrogen, GCS = Glasgow Coma Scale, CRP = serum C-reactive protein, LDH = serum lactate dehydrogenase, Resp = respiratory, $\mathrm{Spo}_2$ = oxygen saturation.

maximum score for the full encounter. For encounters who did develop ARF, we calculated the maximum score up to 1 hour before the onset of respiratory failure. We used the maximum scores for all encounters to compute the ROC AUC. The area under the curve (AUC) point estimates and SEs were computed using bootstrap resampling of encounters (after the models have been fit). We also calculated the area under the precision recall curve (PR AUC) for the models using multiple markers. To compare with an existing model, we calculated the ROC AUC and PR AUC using the maximum Modified Early Warning Score (MEWS) for each encounter (18–21). We used the Python shap package version 0.35.0 to estimate the effect size of predictors (22), which allows us to estimate the importance of individual features for each prediction.

To assess timing of alerts relative to ARF and how the models would behave in a clinical context, we chose a score threshold that correctly excluded 75% of the encounters that did not develop ARF (i.e., a 75% specificity). For all true-positive encounters (defined as encounters where a given model score has a maximum score above the 75% specificity threshold) without an alert on admit, we calculated the "alert onset" as the first time at which the score crosses the threshold. This method did not require knowledge of when the maximum risk score was reached during an encounter, thus only used data collected prior to alert onset time.

## RESULTS

### Study Population and Demographics

The Mid-Atlantic and UW cohorts are summarized in **Table 1**. After applying exclusions, our study includes 1,741 encounters from the Mid-Atlantic system and 274 from the UW health system (**Fig. S1**, http://links.lww.com/CCX/A635). Rate of ARF is comparable across the two systems. The median time from admission to onset of ARF is 33 hours in the Mid-Atlantic cohort and 29 hours in the UW cohort. Mortality rate is 6% and 4% at the Mid-Atlantic health system and UW,respectively, with fewer than 5% of patients still in the hospital at the end of the follow-up period. The median length of stay at the Mid-Atlantic system is approximately 4.5 days but approximately 3.5 days at UW. Among excluded cases for surgical procedures, 107 of 118 (90.6%) in the Mid-Atlantic cohort and 31 of 33 (93.9%) in the UW cohort were clearly unrelated to COVID-19 (**Tables S4**, and **S5**, http://links.lww.com/CCX/A635).

### Prognostic Value of Markers and Trajectory Data

The marker-specific AUC using the full models built on clinical trajectory data is shown in **Figure 1*A***. In **Figure 1*B***, we plot the AUC of these models against the AUC of the latest value models for each marker. For markers with an AUC of 0.60 or greater, models incorporating clinical data trajectory features outperform models that only use the latest value of the marker.

### Assessing Marker Alert Timing Relative to ARF

Eight markers have an AUC greater than 0.60 and are included in the timeliness assessment. For these markers, the time between alert onset and escalation of respiratory failure is summarized in **Figure 1*C***. Markers of inflammatory response (temperature, C-reactive protein [CRP], lactate dehydrogenase [LDH]) yielded an alert earlier in the hospital course than markers of respiratory compromise (respiratory rate, oxygen saturation [$\text{Spo}_2$], arterial blood gas measurements).

Among all true-positive encounters with an alert from at least one marker, we found 83% of encounters with an inflammation marker alert (temperature, LDH, or CRP) also had a respiratory marker alert ($\text{Spo}_2$ or respiratory rate), whereas only 40% of encounters with a respiratory marker alert also had an inflammation marker alert. Of the encounters with only a respiratory marker alert, 57% had at least one measurement of LDH or CRP. In encounters with both an inflammation and respiratory marker alert, the inflammation alert preceded the respiratory alert in over 75% of the encounters.

### Comparison to ARF in Other Types of Pneumonia

Models for COVID and non–COVID pneumonia demonstrated similar abilities to predict ARF in the clinical cohorts in which they were derived (ROC AUC 0.80 vs ROC AUC 0.85 for COVID and pneumonia cohorts, respectively) (**Fig. 2, *A*** and ***B***). However, they demonstrated poor performance when applied out of this context when cross-validated against one another

## TABLE 1.
Patient Population

| Characteristics and Outcomes | Mid-Atlantic | University of Washington | Pneumonia |
|---|---|---|---|
| Total, *n* | 1,741 | 274 | 3,475 |
| Median age (IQR) | 58 (42–72) | 58 (41–69) | 68 (54–80) |
| Female, % (*n*) | 50 (867) | 46 (125) | 52 (1,803) |
| Acute respiratory failure[a], % (*n*) | 13 (233) | 9 (26) | 4 (138) |
| Ventilator[b] | 5 (89) | 4 (11) | 2 (62) |
| Noninvasive positive pressure ventilation[b] | 1 (16) | 3 (9) | 2 (76) |
| ≥ 15 L/min $O_2$ for ≥ 8 consecutive hours[b] | 7 (128) | 2 (6) | Not applicable |
| Median time from admit to acute respiratory failure (IQR) | 1 d 9 hr<br>(13 hr to 3 d 6 hr) | 1 d 5 hr<br>(9 hr to 3 d 9 hr) | 1 d 10 hr<br>(11 hr to 1 d 14 hr) |
| Discharged, % (*n*) | 90 (1,561) | 93 (255) | 98 (3,416) |
| Died, % (*n*) | 6 (101) | 4 (10) | 2 (59) |
| Median length of stay for discharged/died (IQR) | 4 d 11 hr<br>(2 d 4 hr to 7 d 22 hr) | 3 d 16 hr<br>(1 d 18 hr to 8 d 16 hr) | 2 d 22 hr<br>(1 d 19 hr to 4 d 21 hr) |

IQR = interquartile range.

[a] Number of patients with acute respiratory failure after exclusion of patients (per study protocol) meeting the outcome definition while still in the emergency department or those admitted directly to ICU. The occurrence rate of intubation in the overall hospitalized populations was 15% and 18% in the Mid-Atlantic and University of Washington cohorts, respectively.

[b] Each patient was only counted once based on the maximal intervention received (ventilator > noninvasive positive pressure ventilation > 15 L/min $O_2$).

(AUC 0.68 for COVID patients on pneumonia model and AUC 0.75 for pneumonia patients on COVID model). A limited number of markers were found to discriminate these disease-specific models (**Fig. 2*C***), with $Spo_2$ and creatinine demonstrating the greatest difference in prediction of COVID vs other types of pneumonia.

### Prognostic Value of Combining Markers

In **Figure 3*A***, we plot the ROC AUC and PR AUC for ARC on the pooled Mid-Atlantic and UW validation cohorts. The ROC AUC for ARC is 0.80 and PR AUC is 0.36. On the same pooled validation cohorts, the MEWS has an ROC AUC of 0.65 and PR AUC of 0.18. In Figure 3*B*, we show ARC's AUC on the Mid-Atlantic and UW cohorts separately, further broken down by the model's ability to discriminate patients on varying levels of supplemental oxygen.

To characterize the clinical potential of ARC, we performed a quasiprospective evaluation using a model score threshold with 75% specificity to calculate the first time at which the model score crosses that threshold. Using this threshold, ARC alerts on 58 of 75 positive encounters in the pooled validation set at least one hour before ARF onset (77% sensitivity). At the overall event rate of 10% in our cohort, 29% of patients with a high risk designation progressed to ARF versus 4% without. Forty-eight of 58 true-positive alerts (83%) occur on the floor or in observation units (i.e., prior to ICU admission), and the median time from alert to ARF onset is 32 hours (interquartile range, 16–63 hr). The timing of key events for all true positives in the pooled Mid-Atlantic and UW validation cohorts is shown in **Figure 4*A*** and summarized in **Table S7** (http://links.lww.com/CCX/A635). **Figure 4*B*** shows the progression of an example encounter, which is addressed in the discussion. The features contributing to the ARC score at the time of the alert are shown in **Figure 4*C***.

Many false-positive alerts fire on patients with moderate supplemental oxygen requirements but who fall short of meeting the ARF criteria. On the left-most side of the *x*-axis in Figure 3*B*, we show ARC's AUC on the Mid-Atlantic and UW cohorts when discriminating between patients on no supplemental oxygen and patients who develop ARF. ARC has excellent discrimination in

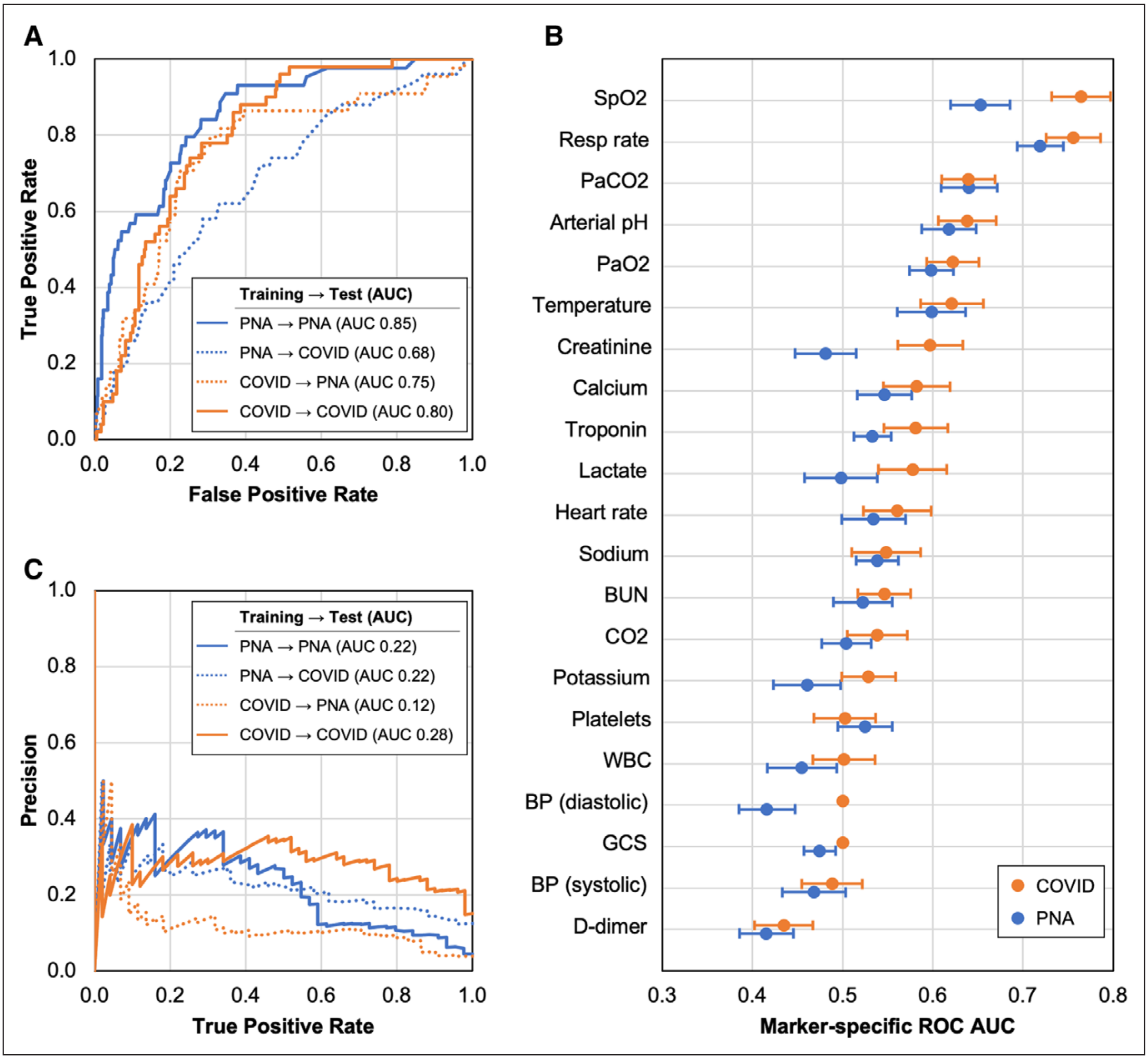

**Figure 2.** Comparison of acute respiratory failure (ARF) in coronavirus disease 2019 (COVID-19) patients with other types of pneumonia. **A**, Receiver operating characteristic (ROC) and (**B**) precision recall curves for models trained on Mid-Atlantic COVID-19 and pneumonia cohorts. Area under the curve (AUC) for each curve is shown in the inset of each plot. **C**, Discriminative power of individual markers in models trained on COVID-19 (*orange*) or pneumonia (*blue*) populations and tested against Mid-Atlantic COVID-19 cohort. AUC point estimates and SEs were computed using bootstrap resampling of encounters. BP = blood pressure, BUN = blood urea nitrogen, GCS = Glasgow Coma Scale, PNA = pneumonia, $Spo_2$ = oxygen saturation.

this case. As we include patients with increasing supplemental oxygen requirements, the AUC declines.

We used the Python shap package (22, 23) to estimate the importance of individual features for each prediction. The distribution over estimated importances across all test set predictions for the 10 most important features is shown in **Figure 3*C***.

## DISCUSSION

We present the development of a specific prognostic model for the risk of ARF due to COVID-19 among patients admitted to non-ICU inpatient environments, as well as its validation in an independent cohort of patients from a second medical system. We further

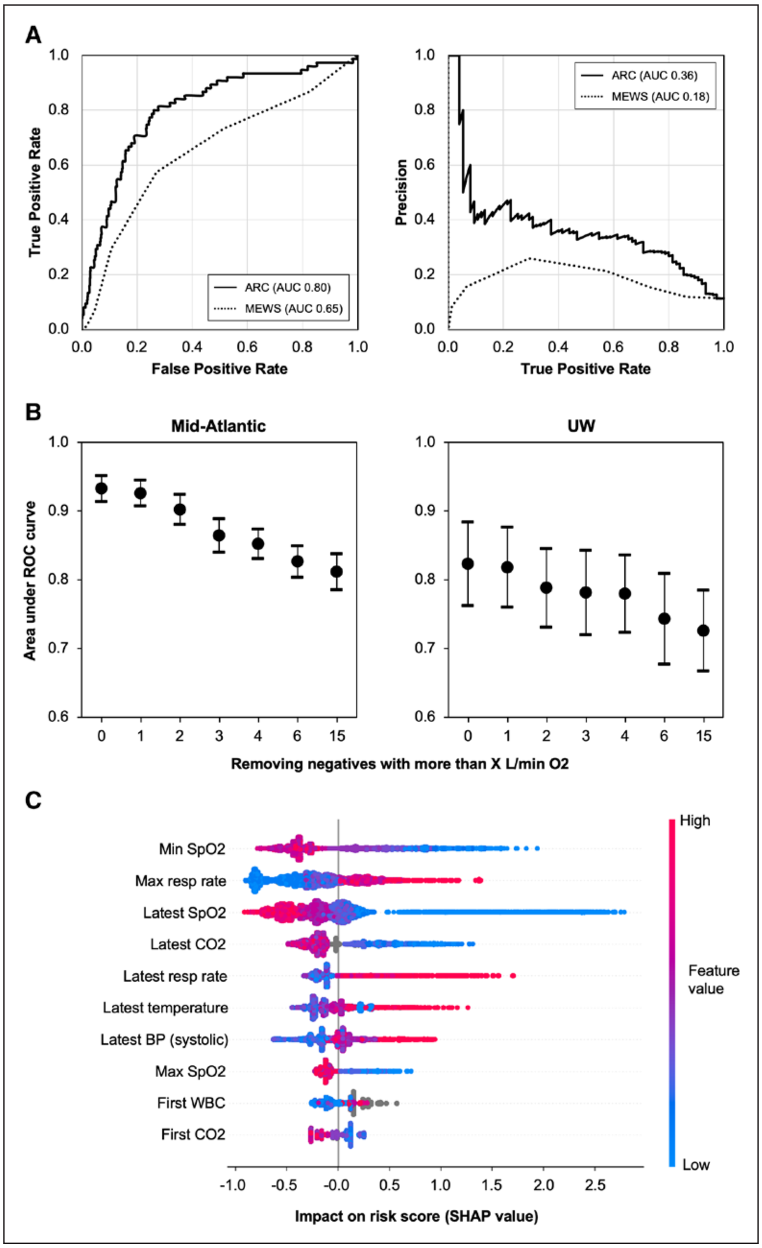

**Figure 3.** Prognostic value of combining markers. **A**, Receiver operating characteristic (ROC) (*left*) and precision recall (*right*) curves for ARC and Modified Early Warning Score (MEWS) validated on the pooled Mid-Atlantic validation subset and University of Washington (UW) coronavirus disease 2019 (COVID-19) cohorts. **B**, Ability of ARC to. discriminate between increasingly inclusive sets of patients without acute respiratory failure (moving from *left* to *right* on the *x*-axis). The *left*-most measurements display area under the curve (AUC) for each model when discriminating between patients who receive no supplemental oxygen and who develop acute respiratory failure (ARF). The *right*-most points include all patients that do not meet our criteria for ARF. ROC AUC point estimates and SEs are plotted separately for Mid-Atlantic and UW test cohorts. **C**, The distribution of the importance of each marker across the pooled Mid-Atlantic and UW test cohorts is shown for the 10 most important features in the ARC model. High absolute value SHAP score indicates a large relative contribution to the overall model score. ARC = anticipating respiratory failure in coronavirus disease, BP = blood pressure, SHAP = SHapley Additive exPlanations, $Spo_2$ = oxygen saturation.

describe the explanatory features and temporal progression of measures that predict clinical deterioration and compare these with parallel models developed in patients with alternative forms of respiratory infection. The most important variables predicting impending respiratory decline among patients hospitalized with COVID-19 in this study ($Spo_2$, respiratory rate, inflammatory markers) are consistent with previous reports (5–7, 24, 25). The vast majority of true-positive alerts (83%) occurred while patients were in non-ICU environments and occurred long enough before the onset of ARF (median 32 hr) to be clinically meaningful. We found that a COVID-19–specific model yielded improved performance over a general model of respiratory failure in pneumonia or an existing clinical deterioration score (MEWS). The potential clinical utility of the model is exemplified by the timeline of postadmission events shown for an individual patient in Figure 4*B*. At admission, this patient has unremarkable oxygen saturation and respiratory rate. Initially on minimal oxygen supplementation, after 2 days, the patient experienced a mild desaturation which prompted a temporary increase in the flow rate to above 2 L/min. Shortly thereafter, the risk score rose above the alert threshold indicating an increasing probability of acute respiratory decline despite the fact that no single measure reached a range that would prompt clinical concern for several more hours. Approximately 24 hours after the alert, oxygen therapy was rapidly stepped up over the course of several hours before transfer to the ICU and intubation. To make the alert more interpretable in a clinical deployment, displaying explanations of the importance of individual features (as shown in Fig. 4*C*) would provide additional context to the ARC score to supplement clinician judgment.

## Importance of Trajectory Data and Specific Markers

Models harnessing data about patient trajectory yielded improved performance over those examining only a snapshot. The history of $Spo_2$ and respiratory rate (Fig. 1*A*) provided important additional power to predict risk of progression and, matching clinical intuition, was considerably more discriminative than other markers. Incorporating trajectory information likely offers benefits over institutional triage criteria that traditionally rely on instantaneous measures of clinical status. For example, criteria for triggering "rapid response" evaluations of patient status or ICU admission at both study centers use static thresholds based on oxygen delivery and saturation at the time of evaluation, consistent with general practice COVID-19 guidelines (26). Thoughtfully calibrated automated alerts based on trajectory data may improve the consistency with which these routinely generated clinical data are used while helping busy and potentially overwhelmed care teams to identify patients at the highest risk of decompensation.

We observed that the predictive nature of specific marker classes evolved in a stereotyped fashion over the course of inpatient admission. Inflammatory markers (e.g., temperature, LDH, CRP) were found to precede respiratory markers ($\mathrm{Spo}_2$ and respiratory rate) in their ability to identify impending respiratory failure. Although identified as an important risk factor for death and composite outcomes in some of the first COVID-19 studies (24, 25), D-dimer was not predictive of ARF in the COVID-19 cohorts studied here.

## Applicability and Comparison With Other Pneumonias

There is active debate regarding the relationship between ARF in other forms of pneumonia and in COVID-19; for example, Gattinoni et al (9) argue that ARF in COVID-19 patients is physiologically distinct from that commonly seen in patients who develop acute respiratory distress syndrome as a result of other causes and that it therefore must be treated with a different approach. This position has been controversial (10–12). Our models are consistent with the hypothesis that these entities are physiologically distinct; the SARS-CoV-2–specific prognostic models developed here performed better within the SARS-CoV-2 cohort than the model developed against patients with other forms of pneumonia. The main difference was in the performance of $\mathrm{Spo}_2$ as a predictor. $\mathrm{Spo}_2$ performed well when our model was trained on the pure COVID cohort but was significantly weaker against the pneumonia cohort. Although not definitive, these findings are consistent with clinical observations of severe hypoxia out of proportion to symptoms in patients with COVID-19 (27), though the findings could also be explained by the use of protocolized respiratory interventions specific to COVID-19.

## Comparison With Prior Studies

Several predictive models related to COVID-19 have been published. Many existing prognostic models for COVID-19 are designed to make predictions at time of admission (5–7) or do not explicitly use information about the history of important clinical markers into account (28). The READY trial, which evaluated data from early in admission to predict need for invasive mechanical ventilation, is only applicable to differentiating very sick patients that may not be difficult to identify clinically (29). A study examining the outcome of mortality generated interest (28), although the clinical utility of that model has been questioned on the basis of failure of the model to be externally validated (30). Although the ISARIC 4C model benefited from a very large multi-dataset, it was also designed to make a prediction only at the time of admission (8). In contrast to prior studies, we found that continuous incorporation of routinely collected information over the course of an inpatient encounter enabled enhanced prediction of clinical deterioration within an actionable time window of approximately 24 hours. Among cases correctly identified as progressing to respiratory failure by ARC, 72% (42/58) had risk scores below the threshold near the time of admission (Fig. 4*A*).

## Strengths and Limitations

A strength of our work is the inclusion of data representing eight hospitals from two distinct health systems with different practice patterns and experience with respect to the surge of COVID-19 patients during the initial months of the pandemic (UW experiencing an earlier, shallower peak compared with the Mid-Atlantic region). The modeling approaches used in the present work are standard techniques in the field of machine learning, thus mitigating risk associated with novel or esoteric analytical techniques. Monitoring throughout an encounter is a strength of ARC over other models that make predictions only at the time of admission. However, we still observed wide variation in time between first alert and onset of ARF. Additional refinements and an alternative selection of operating variables could make ARC more actionable for clinical teams. In addition to the limitations common to all retrospective analyses (i.e., unobserved confounders, e.g., because the patients were already under suspicion

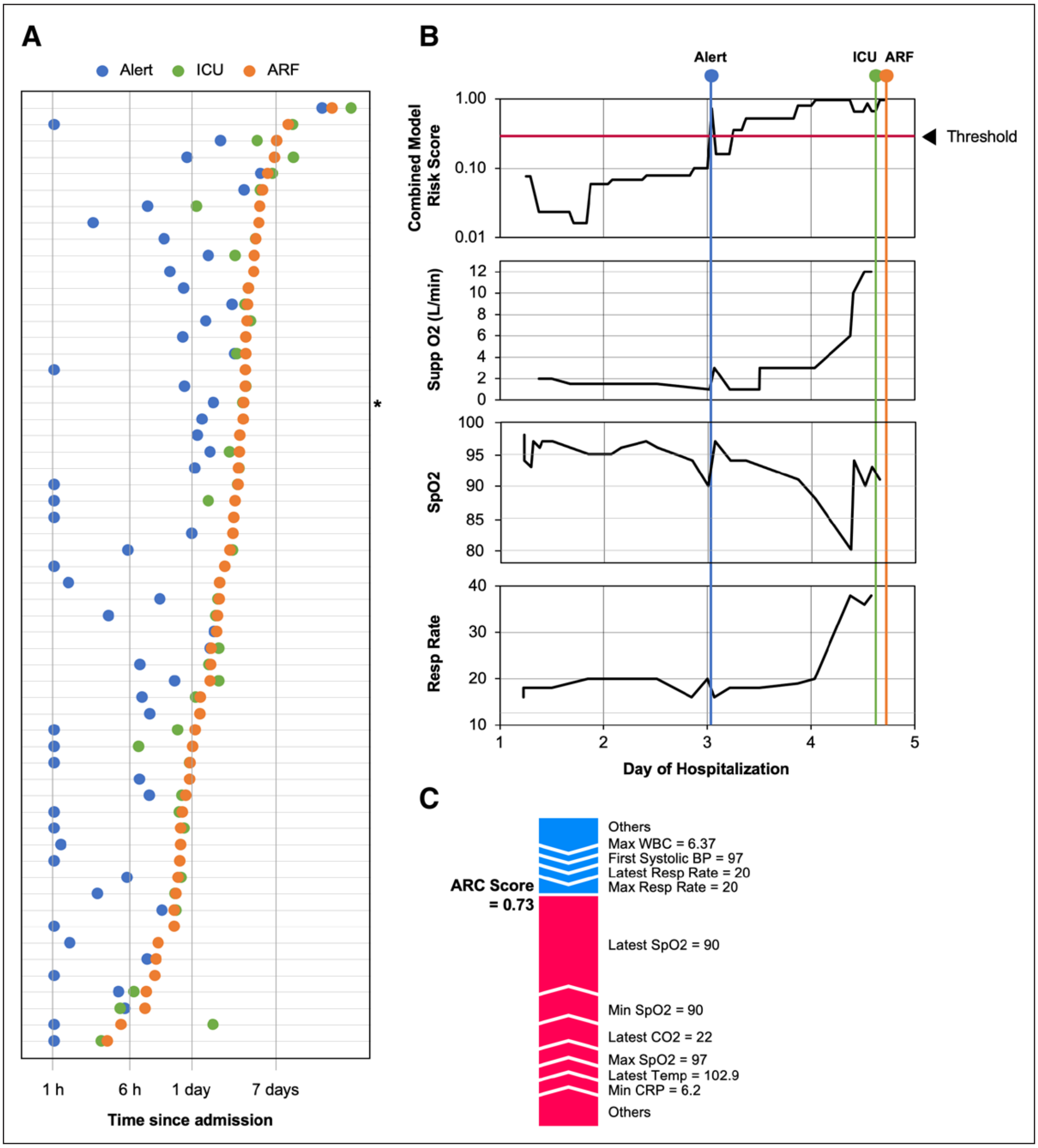

**Figure 4.** Timing of key events. **A**, Timing of events relative to time of admission for all true-positive encounters identified by the ARC model on the pooled Mid-Atlantic validation subset and University of Washington coronavirus disease 2019 cohorts. Each row represents a single encounter. Times between admission to an inpatient floor and the event are plotted on a logarithmic scale. *Blue*: time since admission ARC score first crosses a threshold selected for 75% specificity. Alert onset times within the first hour of admission are plotted at 1 hr to improve visualization. *Green*: time patient is first admitted to an ICU setting. *Orange*: onset of acute respiratory failure (ARF). The asterisk (*) indicates the encounter highlighted in (**B**). **B**, Progression of an example patient in the Mid-Atlantic cohort leading to intubation on the fourth day of hospitalization. *Top* panel: ARC score recalculated at every hour since admission, with alert threshold shown as *red horizontal line*. Second panel: the amount of supplemental oxygen ($O_2$) that the patient receives; the time when the patient is intubated is shown on the same panel. Third panel: Patient's $Spo_2$. *Bottom* panel: Respiratory (Resp) rate. The times when the ARC alert fires (*blue*), the patient is transferred to the ICU (*green*), and the patient meets the criteria for ARF (*orange*) are shown as *vertical lines*. Additional details about the sampling density for supplemental $O_2$, $Spo_2$, and Resp rate are shown in **Table S6** (http://links.lww.com/CCX/A635). **C**, Features with the highest contribution to the ARC score at the time of the alert. Features increasing ARC score are shown in *red*, and features decreasing ARC score are shown in *blue*. The height of each segment indicates the effect of each feature based on the SHapley Additive exPlanations value. ARC = anticipating respiratory failure in coronavirus disease, CRP = C-reactive protein, $Spo_2$ = oxygen saturation.

from clinical teams; bias in underlying data; undetected data quality issues), several specific limitations should be noted. Combining bacterial and viral pneumonia cases in the pneumonia cohort may result in mixing data from two distinct subpopulations of patients. Encounters were included on the basis of a prior positive COVID-19 test result; some admissions may have occurred due to indications other than COVID-19–related care such as necessary surgical procedures. Our approach to screening these admissions (any surgical procedure occurring during the admission) may have resulted in exclusion of a small number of true hospitalizations for COVID-19, which also involved intervening surgical procedures related to complications of COVID-19. Despite using data from multiple centers, the overall number of positive events was small. Compliance with important HIPAA regulations prevented the use of dates of service; given the ongoing evolution of clinical practice as more is learned about COVID-19, there are likely to have been changes to the underlying management of oxygen therapy and respiratory support that may have impacted the analysis in unknown ways.

## CONCLUSIONS

ARC enables prediction of postadmission escalations in the intensity of respiratory support over the entirety of the clinical encounter within a clinically actionable timeframe. The ability to identify elevations in patient risk throughout their hospital trajectory (including patients not drawing clinical suspicion at admission) presents important opportunities for improving outcomes by using anticipatory rather than reactive approaches to escalation of care. In the United States, where most major U.S. hospitals have achieved the highest stages of the Healthcare Information and Management Systems Society electronic medical record (EMR) adoption model, implementation of the proposed solution through already available integration interfaces is feasible. We were able to accomplish this with a major EMR vendor with a week's worth of resources from the hospital's information technology team. Our live implementation was also supplemented with real-time monitoring for stability and robustness to shifts (31).

This anticipatory approach has applications at both the individual patient- and hospital system-levels. Individually, there could be improved alignment of monitoring resources with patient need, targeted deployment of respiratory care resources, advanced planning for potential changes in patient status, or new criteria for early initiation of therapies. For example, current national guidelines are clear on glucocorticoid treatment for mechanically ventilated patients in COVID-19 ARF (strongly recommended) and for patients not on supplemental oxygen (treatment not indicated, strong recommendation) (32). This leaves a large swath of patients for whom no clear recommendation exists. Early indicators of disease progression may be used to guide treatment initiation in this intermediate group. At a system level, the ability to monitor the active, admitted cohort of COVID-19 patients may allow for improved planning around bed flow, staffing, and resource allocation (allowing certain elective procedures requiring postoperative ICU care to proceed, for instance).

In a clinical deployment, pairing risk predictions with explanations of the factors contributing to the risk score would facilitate clinical interpretation of alerts, allowing providers to engage with automated evaluations of patient trajectory in a way that leverages the combined value of both clinical and machine learning–derived assessments. Implementation and evaluation of performance under real-time use in clinical practice remain important future steps to assure generalizable and ongoing utility as the medical response to COVID-19 evolves.

## ACKNOWLEDGMENTS

We would like to thank the Center for Perioperative and Pain initiatives in Quality Safety Outcome and the Department of Anesthesiology and Pain Medicine at the University of Washington, Seattle, for support of this project including provision of expertise to obtain some of the data extract analyzed in the present work. We also thank L. Rosen for additional data analyses.

[1] Computer Science and Statistics, Whiting School of Engineering, Johns Hopkins University, Baltimore, MD.

[2] Health Policy and Management, Bloomberg School of Public Health, Johns Hopkins University, Baltimore, MD.

[3] Bayesian Health, New York, NY.

[4] Critical Care Medicine, University of Pittsburgh Medical Center, Altoona, PA.

[5] Biomedical Informatics and Medical Education, University of Washington, Seattle, WA.

[6] Anesthesiology and Pain Medicine, University of Washington, Seattle, WA.

[7] Pediatric Anesthesiology, Seattle Children's Hospital, Seattle, WA.

*Supplemental digital content is available for this article. Direct URL citations appear in the printed text and are provided in the HTML and PDF versions of this article on the journal's website (http://journals.lww.com/ccejournal).*

*Drs. Saria and Schulam contributed equally. Drs. Sunshine, Long, and O'Reilly-Shah contributed equally.*

*Supported, in part, by National Institutes of Health/National Institute of General Medical Sciences grant ID T32 GM086270-11 (to Dr. Long), National Institutes of Health grant ID K23DA046686 (to Dr. Sunshine), and National Science Foundation grant IDs 1914873 and 1812559 (to Dr. Sunshine).*

*Dr. Saria is a founder of, and holds equity in, Bayesian Health. Dr. Saria also holds equity in and is a member of the scientific advisory board for PatientPing. This arrangement has been reviewed and approved by Johns Hopkins University in accordance with its conflict-of-interest policies. The remaining authors have disclosed that they do not have any potential conflicts of interest.*

*For information regarding this article, E-mail: ssaria@cs.jhu.edu; voreill@uw.edu*

## REFERENCES

1. Thevarajan I, Buising KL, Cowie BC: Clinical presentation and management of COVID-19. *Med J Aust* 2020; 213:134–139
2. Zhang Y, Xiao M, Zhang S, et al: Coagulopathy and antiphospholipid antibodies in patients with COVID-19. *N Engl J Med* 2020; 382:e38
3. Gupta S, Batt J, Bourbeau J, et al: Triaging access to critical care resources in patients with chronic respiratory diseases in the event of a major COVID-19 surge: Key highlights from the Canadian Thoracic Society (CTS) position statement. *Chest* 2020; 158:2270–2274
4. Wynants L, Van Calster B, Collins GS, et al: Prediction models for diagnosis and prognosis of covid-19 infection: Systematic review and critical appraisal. *BMJ* 2020; 369:m1328
5. Liang W, Liang H, Ou L, et al: 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; 180:1081–1089
6. Vaid A, Somani S, Russak AJ, et al: Machine learning to predict mortality and critical events in a cohort of patients with COVID-19 in New York City: Model development and validation. *J Med Internet Res* 2020; 22:e24018
7. Garibaldi BT, Fiksel J, Muschelli J, et al: Patient trajectories among persons hospitalized for COVID-19: A cohort study. *Ann Intern Med* 2021; 174:33–41
8. Gupta RK, Harrison EM, Ho A, et al: Development and validation of the ISARIC 4C Deterioration model for adults hospitalised with COVID-19: A prospective cohort study. *Lancet Respir Med* 2021; 9:349–359
9. Gattinoni L, Chiumello D, Caironi P, et al: COVID-19 pneumonia: Different respiratory treatments for different phenotypes? *Intensive Care Med* 2020; 46:1099–1102
10. Bos LDJ, Paulus F, Vlaar APJ, et al: Subphenotyping acute respiratory distress syndrome in patients with COVID-19: Consequences for ventilator management. *Ann Am Thorac Soc* 2020; 17:1161–1163
11. Jain A, Doyle DJ: Stages or phenotypes? A critical look at COVID-19 pathophysiology. *Intensive Care Med* 2020; 46:1494–1495
12. Rajendram R, Kharal GA, Mahmood N, et al: Rethinking the respiratory paradigm of COVID-19: A "hole"in the argument. *Intensive Care Med* 2020; 46:1496–1497
13. Collins GS, Moons KGM: Reporting of artificial intelligence prediction models. *Lancet* 2019; 393:1577–1579
14. Norgeot B, Quer G, Beaulieu-Jones BK, et al: Minimum information about clinical artificial intelligence modeling: The MI-CLAIM checklist. *Nat Med* 2020; 26:1320–1324
15. World Health Organization: *WHO R&D Blueprint Novel Coronavirus: COVID-19 Therapeutic Trial Synopsis*. Geneva, Switzerland, World Health Organization, 2020. Available at: https://www.who.int/publications/i/item/covid-19-therapeutic-trial-synopsis
16. Drake MG: High-flow nasal cannula oxygen in adults: An evidence-based assessment. *Ann Am Thorac Soc* 2018; 15:145–155
17. 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*. New York, NY, Association for Computing Machinery, 2016, pp 785–794
18. Subbe CP, Kruger M, Rutherford P, et al: Validation of a modified Early Warning Score in medical admissions. *QJM* 2001; 94:521–526
19. McNarry AF, Goldhill DR: Simple bedside assessment of level of consciousness: Comparison of two simple assessment scales with the glasgow coma scale. *Anaesthesia* 2004; 59:34–37
20. Romanelli D, Farrell MW: AVPU score. *In*: *StatPearls*. Treasure Island, FL, StatPearls Publishing, 2020
21. Sam NTH, Toan PN, Hong TTM, et al: Comparison of AVPU scale and the Glasgow coma scale score in assessing encephalitis in children. *Pediatr Infect Dis* 2016; 1:29
22. Lundberg SM, Lee SI: A unified approach to interpreting model predictions. *In:* Guyon I, Luxburg UV, Bengio S, et al (Eds). *Advances in Neural Information Processing Systems 30*. Red Hook, NY, Curran Associates, 2017, pp 4765–4774
23. Lundberg SM, Nair B, Vavilala MS, et al: Explainable machine-learning predictions for the prevention of hypoxaemia during surgery. *Nat Biomed Eng* 2018; 2:749–760
24. Wu C, Chen X, Cai Y, et al: Risk factors associated with acute respiratory distress syndrome and death in patients with coronavirus disease 2019 Pneumonia in Wuhan, China. *JAMA Intern Med* 2020; 180:934–943
25. Zhou F, Yu T, Du R, et al: Clinical course and risk factors for mortality of adult inpatients with COVID-19 in Wuhan, China: A retrospective cohort study. *Lancet* 2020; 395:1054–1062
26. Liu S, Sweeney C, Srisarajivakul-Klein N, et al: Evolving oxygenation management reasoning in COVID-19. *Diagnosis (Berl)* 2020; 7:381–383
27. Shenoy N, Luchtel R, Gulani P: Considerations for target oxygen saturation in COVID-19 patients: Are we under-shooting? *BMC Med* 2020; 18:260

28. Yan L, Zhang HT, Goncalves J, et al: An interpretable mortality prediction model for COVID-19 patients. *Nature Machine Intelligence* 2020; 2:283–288

29. Burdick H, Lam C, Mataraso S, et al: Prediction of respiratory decompensation in Covid-19 patients using machine learning: The READY trial. *Comput Biol Med* 2020; 124:103949

30. Barish M, Bolourani S, Lau LF, et al: External validation demonstrates limited clinical utility of the interpretable mortality prediction model for patients with COVID-19. *Nature Machine Intelligence* 2021; 3:25–27

31. Subbaswamy A, Saria S: From development to deployment: Dataset shift, causality, and shift-stable models in health AI. *Biostatistics* 2020; 21:345–352

32. COVID-19 Treatment Guidelines: Corticosteroids. 2020. Available at: https://www.covid19treatmentguidelines.nih.gov/immune-based-therapy/immunomodulators/corticosteroids/. Accessed November 3, 2020