OPEN

Prediction of Acute Respiratory Failure Requiring Advanced Respiratory Support in Advance of Interventions and Treatment: A Multivariable Prediction Model From Electronic Medical Record Data

An-Kwok I. Wong, MD, PhD[1]
Rishikesan Kamaleswaran, PhD[2]
Azade Tabaie, MS[2]
Matthew A. Reyna, PhD[2]
Christopher Josef, MD[2]
Chad Robichaux, MS[2]
Anne A. H. de Hond, MSc[3,4]
Ewout W. Steyerberg, PhD[3]
Andre L. Holder, MD, MSc[1]
Shamim Nemati, PhD[5]
Timothy G. Buchman, MD, PhD[6]
James M. Blum, MD[2,7]

BACKGROUND: Acute respiratory failure occurs frequently in hospitalized patients and often begins outside the ICU, associated with increased length of stay, cost, and mortality. Delays in decompensation recognition are associated with worse outcomes.

OBJECTIVES: The objective of this study is to predict acute respiratory failure requiring any advanced respiratory support (including noninvasive ventilation). With the advent of the coronavirus disease pandemic, concern regarding acute respiratory failure has increased.

DERIVATION COHORT: All admission encounters from January 2014 to June 2017 from three hospitals in the Emory Healthcare network (82,699).

VALIDATION COHORT: External validation cohort: all admission encounters from January 2014 to June 2017 from a fourth hospital in the Emory Healthcare network (40,143). Temporal validation cohort: all admission encounters from February to April 2020 from four hospitals in the Emory Healthcare network coronavirus disease tested (2,564) and coronavirus disease positive (389).

PREDICTION MODEL: All admission encounters had vital signs, laboratory, and demographic data extracted. Exclusion criteria included invasive mechanical ventilation started within the operating room or advanced respiratory support within the first 8 hours of admission. Encounters were discretized into hour intervals from 8 hours after admission to discharge or advanced respiratory support initiation and binary labeled for advanced respiratory support. Prediction of Acute Respiratory Failure requiring advanced respiratory support in Advance of Interventions and Treatment, our eXtreme Gradient Boosting-based algorithm, was compared against Modified Early Warning Score.

RESULTS: Prediction of Acute Respiratory Failure requiring advanced respiratory support in Advance of Interventions and Treatment had significantly better discrimination than Modified Early Warning Score (area under the receiver operating characteristic curve 0.85 vs 0.57 [test], 0.84 vs 0.61 [external validation]). Prediction of Acute Respiratory Failure requiring advanced respiratory support in Advance of Interventions and Treatment maintained a positive predictive value (0.31–0.21) similar to that of Modified Early Warning Score greater than 4 (0.29–0.25) while identifying 6.62 (validation) to 9.58 (test) times more true positives.

DOI: 10.1097/CCE.0000000000000402

Furthermore, Prediction of Acute Respiratory Failure requiring advanced respiratory support in Advance of Interventions and Treatment performed more effectively in temporal validation (area under the receiver operating characteristic curve 0.86 [coronavirus disease tested], 0.93 [coronavirus disease positive]), while achieving identifying 4.25–4.51× more true positives.

CONCLUSIONS: Prediction of Acute Respiratory Failure requiring advanced respiratory support in Advance of Interventions and Treatment is more effective than Modified Early Warning Score in predicting respiratory failure requiring advanced respiratory support at external validation and in coronavirus disease 2019 patients. Silent prospective validation necessary before local deployment.

KEY WORDS: acute respiratory failure; data mining; early warning scores; electronic health records; machine learning; prediction

Acute respiratory failure (ARF)—a disorder characterized by functional lung impairment resulting in hypoxemia, hypercapnia, or both—occurs frequently in hospitalized patients and often begins outside of the ICU, increasing length of stay, cost, and mortality (1–3). Delays in decompensation recognition increase the cost of care and lead to worse outcomes (4, 5). ARF is a common cause of critical illness. This is especially true in the pandemic in the coronavirus disease 2019 (COVID-19) era, where ARF is the cardinal sign of critical illness in COVID-19 patients. Fifteen to 20 percent of COVID-19 cases are hospitalized, with 3–5% requiring critical care (6–8). Once requiring mechanical ventilation (MV), initial studies suggested mortalities as high as 30–97%, far higher than other diseases like H1N1 (6–9).

Most current early warning systems, including Modified Early Warning Score (MEWS) and National Early Warning Score, focus on predicting ICU admission, cardiac arrest, or death (10–15). While cardiac arrest and death may be late signs of morbidity and mortality, ICU admission can often be provider dependent, institution dependent, and/or situation dependent (e.g., if an ICU is full) (16). Furthermore, the decision specifically to initiate invasive MV (IMV) is provider dependent—however, there are many forms of advanced respiratory support (AdvRS) a patient with ARF can use, from IMV to less invasive variants of noninvasive ventilation (NIV) and heated humidified high-flow (HHHF) nasal cannula (e.g., Airvo, Fisher Paykel, Irvine, CA, Optiflow, Fisher Paykel, Irvine, CA). Even if a patient has care limitation orders prohibiting intubation, they may be placed on NIV or HHHF. Consequently, the decision for initiation of AdvRS can be seen as a combination of underlying patient physiology and provider practice. For the purpose of this article, any type of AdvRS—including NIV, HHHF, and IMV—that consists of an oxygen source and another device) is often seen as an indicator of increased patient acuity and is a common cause of ICU admission. AdvRS can be viewed as a marker of ARF and a more direct endpoint of respiratory decompensation than IMV or ICU admission.

Prior studies, such as Dziadzko et al (17), have explored the prediction of prolonged IMV (IMV > 48 hr), looking 48 hours in the future. While this is a useful endpoint in care delivery, 48 hours as a threshold for monitoring patients on the floor seems to have less utility. Therefore, we decided to create an algorithm to predict respiratory decompensation within a provider's shift.

The objective of this study is to predict ARF, defined as the need for either invasive or noninvasive AdvRS, and analyze and validate the performance of our algorithm, Prediction of Acute Respiratory Failure requiring advanced respiratory support in Advance of Interventions and Treatment (PARFAIT), on its prediction. We also validate PARFAIT on temporally separate data on patients tested for COVID-19, as COVID was an unexpected respiratory pandemic with unforeseen effects on patient care practice. Notably, Emory Healthcare had provided institutional guidance that early intubation was preferred, HHHF allowed, and NIV to be avoided in ARF in COVID+ patients—a significant departure from previous practice.

METHODS

Protocol design was performed in accordance with the Development and Reporting of Prediction Models by Leisman et al (18), REporting of studies Conducted using Observational Routinely-collected health Data (19), and transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (20) guidelines as a retrospective cross-sectional study. This study was approved by the Emory University Institutional Review Board (No. 33069).

DATA COLLECTION

All adult (> 18 yr old) admission encounters (including obstetrical patients, psychiatric patients, and incarcerated patients but excluding children) from January 2014 to June 2017 and February 2020 to April 2020 from four university-affiliated hospitals in and around Atlanta, GA, had vital signs, laboratory values, oxygen therapy, and demographics extracted. Exclusion criteria included IMV within an operating room (OR) during a single hospital admission and meeting AdvRS criteria within the first 8 hours of admission.

Only data that were routinely collected for clinical care were included. Seventy variable types were provided to the prediction model: two demographics, eight vitals, 46 laboratories, and 14 oxygen therapy variables. Twenty-seven oxygen therapy variables pertinent to AdvRS outcomes were solely used for generation and validation of the AdvRS outcome and explicitly not provided to the model for either training or testing. These variables are demonstrated in **Supplemental Table 1** (http://links.lww.com/CCX/A593). A site-specific data dictionary was created for each hospital to unify data elements. For each variable at each site, data were first statistically cleaned by excluding any outliers beyond 5 SDs from the mean. All data were extracted from the clinical data warehouse at Emory Healthcare, which used data from Cerner Powerchart (Cerner, North Kansas City, MI). Data were linked by encounter identification number across all data sources.

Encounters were windowed into hourly windows from 8 hours after admission to discharge or on intubation. Death was used to mark the end of a hospitalization; all patients, if deceased, were included until death. Each hourly interval was labeled with binary labels indicating any form of MV in the 8-hour event horizon. If a variable had multiple values within an hourly window, the median value was chosen. Models were provided as single values for each variable during each hourly window.

COVID-19 COHORT IDENTIFICATION

From February 2020 to April 2020 cohort, all patients had received a COVID test. Patients were deemed COVID-19+ if they had at least one COVID+ test result. All COVID tests were polymerase chain reaction (PCR) tests (ABI 7000–7500; Roche cobas 6800; Roche Molecular Systems, Branchburg, NJ). Patients were deemed COVID-19– if they never had a positive test result. This method excludes patients who were COVID– by PCR but clinically treated as if they were COVID+, in addition to patients who only tested positive at an external institution but tested negative in the healthcare system. As the model was trained before COVID, the model does not accept COVID input, and therefore, COVID test results were used solely to select patients in the healthcare system.

LABELING

We define an event horizon as the duration of time before the onset of an event—for example, an 8-hour event horizon asks whether an event of interest will occur 8 hours or less from the current window.

The positive label for AdvRS included continuous NIV, HHHF (e.g., Airvo, Fisher Paykel, Optiflow, Fisher Paykel), and IMV. "Traditional" nasal cannula (1–6 L oxygen flow) and "moderate flow" nasal cannula (6–15 L oxygen flow) was not considered to be a positive AdvRS label. Nocturnal NIV, as charted by respiratory therapy, was not considered to be an AdvRS outcome. If an encounter was labeled positive for any reason, the encounter was terminated at that point, regardless of further AdvRS outcomes, as any further transition between AdvRS endpoints would still require ICU admission. For example, if a patient was put on NIV and was later intubated, the encounter data would be terminated at the initiation of NIV and the patient's intubation would never have been "seen" by the algorithm.

MISSING DATA

Missing data were imputed from normal ranges (if not previously performed for the current admission). Missingness rates are described in electronic medical record (EMR) data density, which is characterized in **Supplemental Figure 2** (http://links.lww.com/CCX/A593).

MODELING IMPLEMENTATION

Patient data from three hospitals from January 2014 to June 2017 were then split into five-fold cross validation sets by encounter ("training data," "testing data"). Complete encounters were assigned to folds; no encounter was split across folds. Given the class

imbalance, each training set was then randomly undersampled to give an even case-control split. Patient data from a fourth hospital from January 2014 to June 2017 was used as external validation ("external validation data"). Finally, data from patients who were tested for COVID (including patients with both positive and negative results) from all four hospitals from February 2020 to April 2020 were then used as temporal validation ("COVID-tested validation data"), with COVID+ patients further stratified for further analysis ("COVID+ validation data"). Models were not retrained for both validations for fair assessment. This diagram is noted in **Supplemental Figure 4** (http://links.lww.com/CCX/A593). Data were segmented for the 8 hours prior to an event. For patients never needing AdvRS, a simulated event time was randomly selected from 8 hours after admission to discharge.

We reduce this prediction problem to a regression task—predicting the probability that a patient will develop our endpoint of AdvRS. PARFAIT uses eXtreme Gradient Boosting, which is a supervised learning method that uses gradient boosting with improved regularization to control for overfitting, thus improving performance (21). It constructs an ensemble method (which combines the hypothesis of many "weak" prediction algorithms) of regression trees that are individually adjusted to create a "strong" classifier. Our algorithm computes a prediction score for every hourly window. Eight sequential hourly predictions are combined with a majority vote, and the prediction is evaluated per encounter. A window was positive with a PARFAIT score greater than 0.50.

MEWS was implemented in accordance with the MEWS guidelines (10). Commonly used thresholds include scores greater than 3–5, so those were tested in this study. These values were recomputed for each 1-hour time window, as demonstrated in (**Supplemental Fig. 3**, http://links.lww.com/CCX/A593).

OUTCOMES

A composite qualifying event was defined as initiation of NIV, HHHF, or IMV for any duration of time. An encounter was terminated at the onset of the qualifying event. NIV for the purpose of nocturnal NIV in obstructive sleep apnea as documented in the EMR was excluded from the outcome. Predictions were generated for all hourly windows. The 8 hourly windows preceding an event were combined using a majority vote, with the final prediction evaluated per encounter. Control events were evaluated by generating a random event time from 8 hours after admission through the hour before discharge and performing the same evaluation. Equivocal results (so, four predicting+ and four predicting–) were deemed incorrect (e.g., false positive or false negative), regardless of ground truth label (**Supplemental Fig. 1**, http://links.lww.com/CCX/A593).

Inhospital death, if it occurred, was used to mark the end of a hospitalization.

MODEL COMPARISONS

Models were compared using area under the receiver operating characteristic curve (AUROC), area under the precision-recall curve (AUPRC), sensitivity (also known as recall), specificity, positive predictive value (PPV, also known as precision), and negative predictive value (NPV). Prediction scores for these calculations were derived using the median of the majority class predictions in each of the hourly windows in the 8-hour event horizon prior to the event. Of note, this also implies that the MEWS scores, due to different thresholds changing voting in the event horizon, will have slightly varying values based on threshold. AUROCs on MEWS were calculated using the median MEWS score of the majority vote.

CALIBRATION METHODS

As many machine learning models may be discriminative but have poor calibration, a calibration adjustment model was trained from each fold of the training data by linear regression (22). All subsequent predictions (in validation sets) for both MEWS and PARFAIT were adjusted by these calibration models generated from the training data. No calibration regression models were readjusted.

Calibration was assessed through a calibration plot with predicted probability on the *x*-axis and the observed proportions on the *y*-axis (23, 24). The plot is characterized by an intercept and calibration slope, which should be approximately 0 and 1 for good calibration. Scaled Brier scores are included in

TABLE 1.
Patient Demographics and Data Characteristics

Characteristics	Hospitals 1–3	External Validation Hospital 4	Temporal Validation COVID-Tested Any Result	Temporal Validation COVID+
Count, *n*	82,699	40,143	2,564	389
Cases, *n* (%)	7,524 (9.1)	2,578 (6.4)	426 (16.6)	104 (26.7)
Gender, *n* (%)				
Male	39,021 (47.2)	18,116 (45.1)	822 (32.1)	181 (46.5)
Ethnicity, *n* (%)				
Hispanic or Latino	2,762 (3.3)	488 (1.2)	52 (2)	8 (2.1)
Non-Hispanic or Latino	73,444 (88.8)	37,360 (93.1)	1,482 (57.8)	285 (73.3)
Not recorded	6,493 (7.9)	2,295 (5.7)	1,030 (40.2)	96 (24.7)
Race, *n* (%)				
African American	24,683 (29.8)	32,302 (80.5)	957 (37.3)	243 (62.5)
Asian	2,521 (3)	237 (0.6)	50 (2)	5 (1.3)
Caucasian	51,821 (62.7)	6,918 (17.2)	576 (22.5)	56 (14.4)
American Indian or Alaskan Native	382 (0.5)	48 (0.1)	4 (0.2)	0 (0)
Native Hawaiian or other Pacific Islander	156 (0.2)	37 (0.1)	1 (0)	0 (0)
Other	3,136 (3.8)	601 (1.5)	976 (38.1)	85 (21.9)
Age at visit	62.75 ± 18.82	58.32 ± 18.26	61.36 ± 17.81	59.82 ± 16.24
Outcomes				
In-hospital mortality, *n* (%)	1,339 (1.6)	516 (1.3)	88 (3.4)	43 (11.1)
Hospice, *n* (%)	3,445 (4.2)	1,469 (3.7)	62 (2.4)	2 (0.5)
Length of stay, hospital (d)	4.91 ± 5.72	4.90 ± 6.41	6.83 ± 8.14	8.89 ± 9.38
Temperature	36.71 ± 0.32	36.61 ± 0.41	36.87 ± 0.35	37.08 ± 0.35
Temperature (highest)	37.30 ± 0.56	37.28 ± 0.57	37.70 ± 0.87	38.33 ± 1.01
SBP	126.98 ± 16.44	131.58 ± 19.00	126.43 ± 14.96	126.32 ± 13.71
SBP (lowest)	105.74 ± 15.80	109.67 ± 17.19	100.91 ± 19.48	97.78 ± 21.15
DBP	69.90 ± 10.12	72.54 ± 10.09	72.99 ± 8.98	73.41 ± 9.36
DBP (lowest)	56.20 ± 12.60	60.63 ± 11.24	57.17 ± 14.15	56.19 ± 15.46
HR	80.42 ± 12.49	82.73 ± 12.95	84.48 ± 12.09	86.04 ± 11.41
HR (highest)	98.73 ± 15.86	99.11 ± 15.97	106.43 ± 19.21	110.16 ± 20.53
O_2 saturation	96.97 ± 1.59	96.67 ± 1.64	96.48 ± 1.97	95.70 ± 2.18
O_2 saturation (lowest)	93.81 ± 2.74	98.82 ± 1.44	89.63 ± 13.05	85.22 ± 17.68
Respiratory rate	18.30 ± 1.47	18.72 ± 1.49	18.85 ± 2.26	20.15 ± 2.95
Respiratory rate (highest)	20.90 ± 3.91	21.27 ± 3.84	25.43 ± 19.71	29.31 ± 19.09
Lactate	1.05 ± 0.32	1.07 ± 0.37	1.25 ± 1.01	1.42 ± 1.20
Lactate (highest)	1.13 ± 0.75	1.17 ± 0.84	1.81 ± 3.01	2.34 ± 3.67
Oxygen therapy, *n* (%)				
Room air	81,464 (98.5)	39,626 (98.7)	1,465 (57.1)	314 (80.7)
Nasal cannula/simple mask (1–6 L)	43,630 (52.8)	19,337 (48.2)	988 (38.5)	267 (68.6)
Moderate flow (6–15 L)	3,877 (4.7)	861 (2.1)	162 (6.3)	65 (16.7)
Nocturnal NIV	4,547 (5.5)	2,235 (5.6)	92 (3.6)	3 (0.8)
NIV	3,838 (4.6)	2,574 (6.4)	126 (4.9)	3 (0.8)
Heated humidified high-flow	1,164 (1.4)	391 (1)	234 (9.1)	87 (22.4)
Intubation	4,953 (6)	1,142 (2.8)	385 (15)	98 (25.2)
Intubation > 48 hr	1,225 (1.5)	576 (1.4)	270 (10.5)	80 (20.6)

COVID = coronavirus disease, DBP = diastolic blood pressure, HR = heart rate, NIV = noninvasive ventilation, SBP = systolic blood pressure. Oxygen therapy was calculated as the number of encounters with that oxygen therapy method at any time during their hospitalization. Numeric data (e.g. age at visit, temperature) presented as mean ± SD.

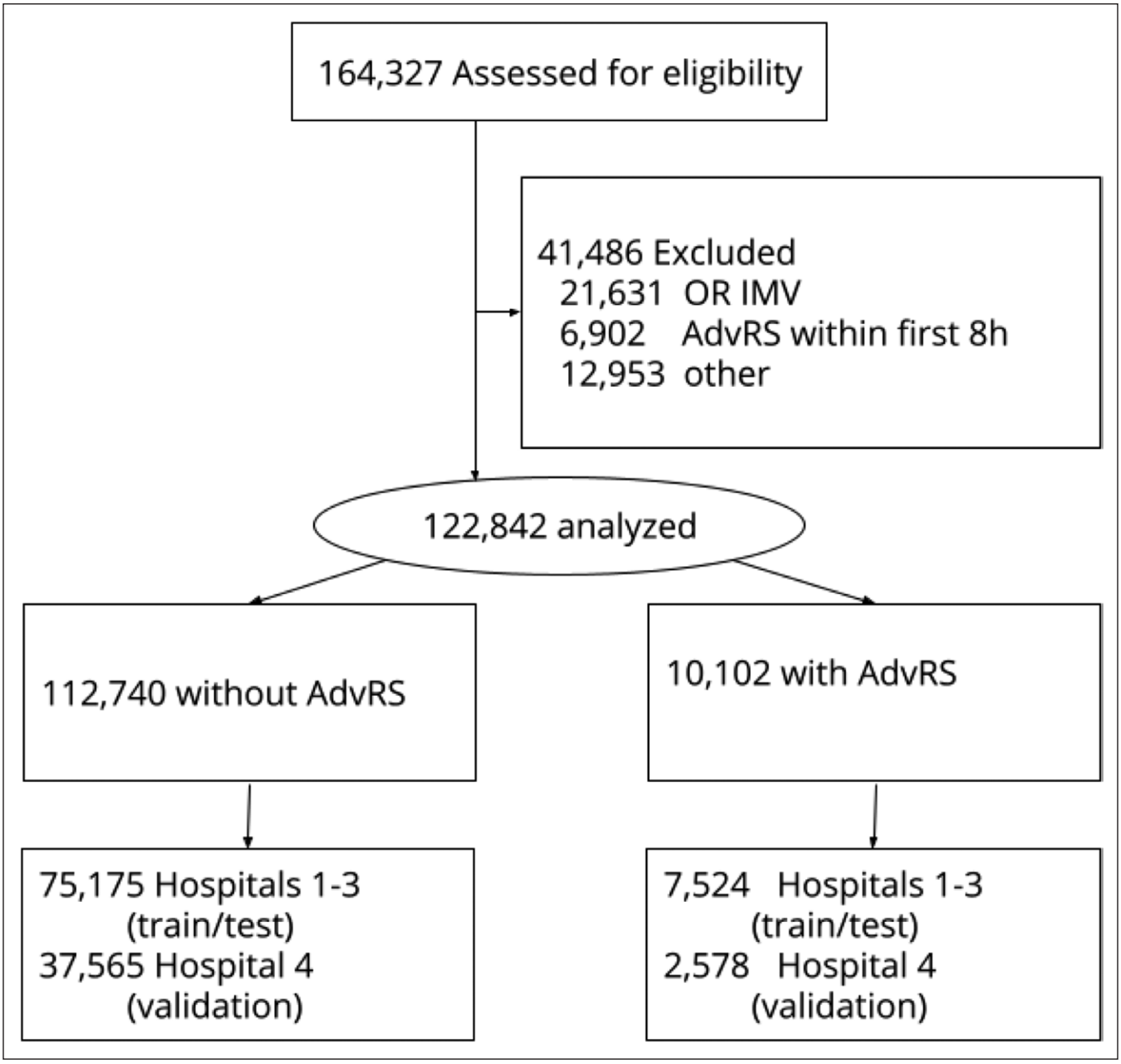

Figure 1. Flow diagram. AdvRS = advanced respiratory support, IMV = invasive mechanical ventilation, OR = operating room.

Supplemental Table 4 (http://links.lww.com/CCX/A593), with equations and methodology from Steyerberg et al (24) further explained in Supplemental Methods (http://links.lww.com/CCX/A593).

RESULTS

Clinical data from January 2014 to June 2017 was extracted from 122,842 encounters, which were expanded into 13,281,322 1-hour windows, with 10,102 cases and 112,740 controls. Patient characteristics are described by dataset in **Table 1** and by flow diagram in **Figure 1**.

Cases and Mortality by Dataset

The training dataset from three hospitals had 75,175 encounters with 7,524 AdvRS cases (9.1%) and 1,332 deaths (1.62%). The external validation dataset from a separate fourth hospital included 40,143 encounters with 2,578 AdvRS cases (6.4%) and 516 deaths (1.29%). The COVID-tested dataset comprised 2,564 encounters from all four hospitals, with 290 AdvRS cases (11.3%) and 88 deaths (3.4%). Finally, the COVID+ dataset is a subset of the COVID-tested dataset, containing 389 COVID+ encounters with 91 AdvRS cases (23.4%) and 43 deaths (11.1%).

Effect of AdvRS on Mortality

The presence of AdvRS was associated with higher mortality in the training dataset (819 [10.9%] vs 520 [0.69%]), the external validation dataset (345 [13.4%] vs 171 [0.5%]), the COVID-tested validation dataset (70 [13.0%] vs 18 [0.89%]), and the COVID+ validation dataset (38 [29.9%] vs 5 [1.9%]).

Effect of AdvRS on Hospice Usage

The presence of AdvRS was associated with higher hospice utilization in the training dataset (682 [9.1%] vs 2,763 [3.68%]), the external validation dataset (341 [13.2%] vs 1,128 [3.0%]), and the COVID-tested validation dataset (26 [4.8%] vs 36 [1.8%]). Hospice usage was significantly reduced in the COVID+ dataset.

Effect of COVID on Clinical Practice

Both COVID-tested and COVID+ validation datasets demonstrate longer lengths of stays than either the training or the external validation dataset (6.83 vs 4.91 d; $p < 0.001$). As expected from a policy in favor of early intubation in COVID+ patients, there is a much higher rate of intubation (385/2,564 [15%] COVID-tested, 98/389 [25.2%] COVID+ vs training dataset 4,953/82,699 [6%]).

PARFAIT MODEL RESULTS

Receiver operating characteristic and precision-recall curves are shown by dataset in **Figure 2**. Model prediction confusion matrices are demonstrated in **Supplemental Table 2** (http://links.lww.com/CCX/A593) and prediction performance is demonstrated in **Supplemental Table 3** (http://links.lww.com/CCX/A593) in the context of varying prevalence in the training, testing, and validation datasets.

PARFAIT Discrimination in Training, Testing, and External Validation Datasets

PARFAIT had significantly better discrimination than MEWS in all datasets as demonstrated by the AUROC (e.g., test dataset PARFAIT AUROC 0.85 vs MEWS AUROC 0.57; DeLong test $p < 0.001$). The PARFAIT misclassification rate was 0.16 on the training dataset and 0.17 on the test dataset. In contrast, the MEWS models with thresholds 3–5 resulted in a misclassification rate of 0.46–0.48 on the training dataset and 0.09–0.15 in the test dataset. Since MEWS was more specific than sensitive, the misclassification rate in MEWS is lower in the test data due to a higher percentage of negative cases.

PARFAIT Sensitivity and PPV in Training, Testing, and External Validation Datasets

At comparable PPVs, PARFAIT (test dataset: 0.31, external validation dataset 0.21) is most similar to MEWS greater than 4 (test dataset: 0.29, external validation: 0.25) while still identifying 6.62-fold (external validation) to 9.58-fold (test) more cases needing AdvRS. As MEWS thresholds increase, stricter criteria lead to lower sensitivity but higher PPV. PPV performance varies significantly between training and test datasets due to the random undersampling used to create the training set.

PARFAIT Discrimination, Sensitivity, and PPV in COVID-Testing Validation Dataset

In the COVID-testing validation dataset, PARFAIT's AUROC (0.86) continues to outperform MEWS' relatively stable AUROCs of (0.66–0.66). Although PARFAIT's PPV (0.36) is lower than that of MEWS (0.42–0.67), PARFAIT's significant improvement in sensitivity (0.84) over MEWS (0.07–0.24) allows it to detect 4.5-fold more cases needing AdvRS.

PARFAIT Discrimination, Sensitivity, and PPV in COVID+ Validation Dataset

The COVID+ validation dataset reinforces the trend toward stronger performance in populations with higher rates of AdvRS, where PARFAIT's AUROC increases to 0.93 (0.92–0.95) in comparison to MEWS (MEWS > 3, 0.73; MEWS > 5, 0.71) with PARFAIT PPV 0.60 similar to MEWS (0.54–0.76) while identifying 4.25–6.86-fold more true positives due to its higher sensitivity (0.94 vs MEWS 0.09–0.29).

PARFAIT Calibration

PARFAIT and MEWS adjusted calibration curves are demonstrated in **Figure 3**, with regressions of the adjusted calibration plot tabulated in Supplemental

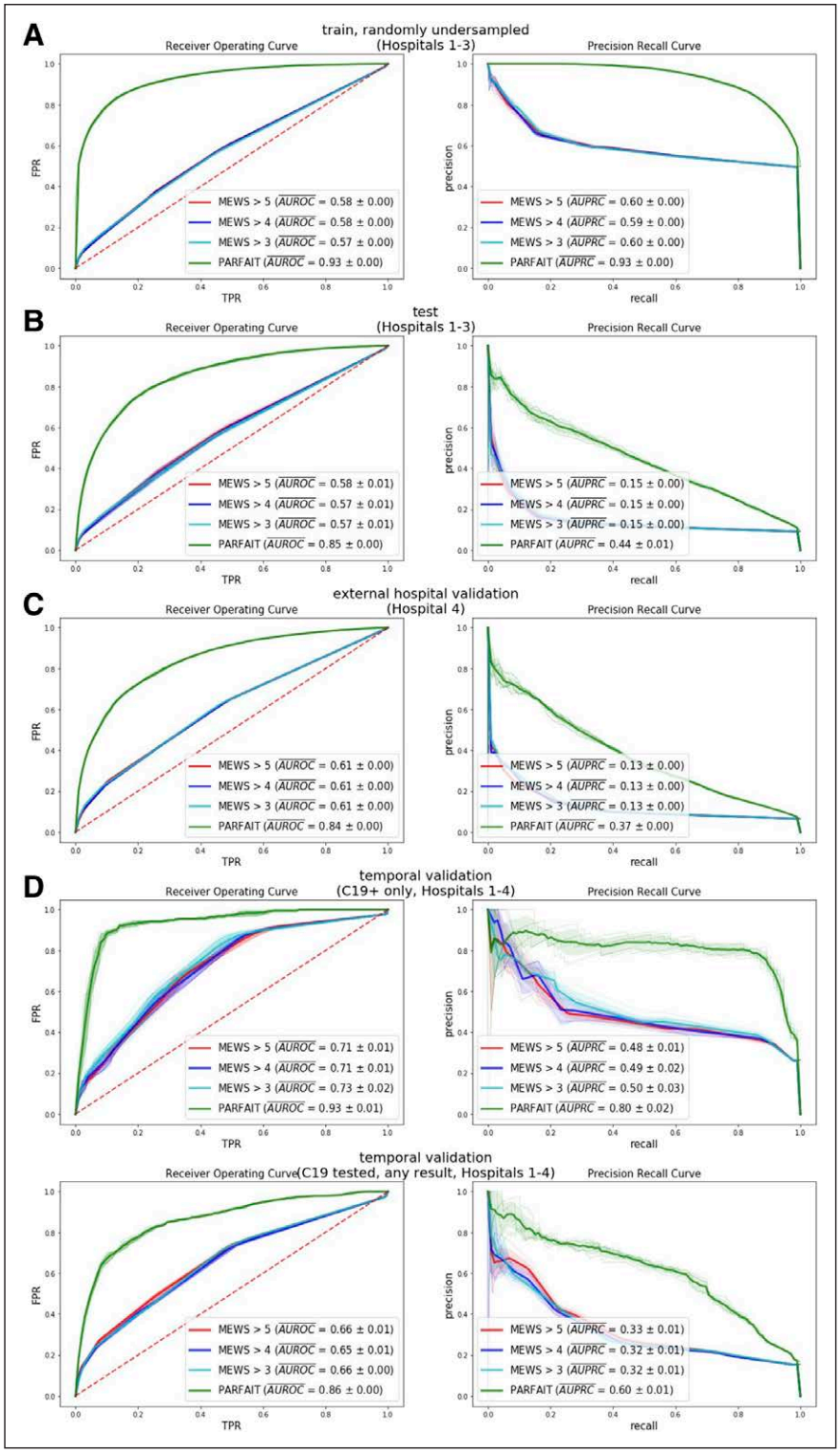

Figure 2. Prediction characterization for PARFAIT and MEWS. Area under the receiver operating characteristic curve (AUROC) and area under the precision-recall curve (AUPRC) plots for PARFAIT versus MEWS (**A**) average train, (**B**) average test, (**C**) average external hospital validation, (**D**) average temporal validation for coronavirus disease 2019 (C19)+ patients, and (**E**) average temporal validation for any patient receiving a C19 test, regardless of result. *Light bands* indicate 1 SD above and below curves. *Dotted plots* indicate individual curves. FPR = false positive rate, MEWS = Modified Early Warning Score, PARFAIT = Prediction of Acute Respiratory Failure requiring advanced respiratory support in Advance of Interventions and Treatment, TPR = true positive rate.

Table 3 (http://links.lww.com/CCX/A593). Scaled Brier scores are shown in Supplemental Table 4 (http://links.lww.com/CCX/A593).

DISCUSSION

We developed and validated an EMR-based risk stratification tool that identifies patients needing AdvRS (NIV, HHHF, IMV) up to 3 hours prior to a patient being placed on support. Five-fold cross-validation and external validation showed that PARFAIT outperformed MEWS at three common thresholds, with a higher PPV, sensitivity, AUROC, and AUPRC. Finally, we demonstrate PARFAIT's robustness within the healthcare system by demonstrating strong performance in temporal validation, even in an unforeseen respiratory pandemic.

Some groups are investigating the utility of predictions for prolonged IMV. Gong et al (25) predicts IMV greater than 48 hours or death 48 hours into the future using random forests, a machine learning technique, with AUROCs 0.77–0.80, false positive rate 0.08–0.17, and PPV 0.13–0.21 (17). However, prolonged IMV is a late endpoint—to reduce the effect of practice variation—and includes a composite endpoint of prolonged IMV and all-cause mortality. All-cause mortality can lead to a less focused endpoint, reducing PPV. Parreco et al (26) predicts the need for IMV greater than 168 hours or tracheostomy for patients receiving IMV in the ICU using gradient boosted trees. However, this focuses on ICU patients already receiving IMV but does not consider predictions on floor patients. Martín-González et al (27) uses multiple machine learning methods to predict NIV success or failure, but this work

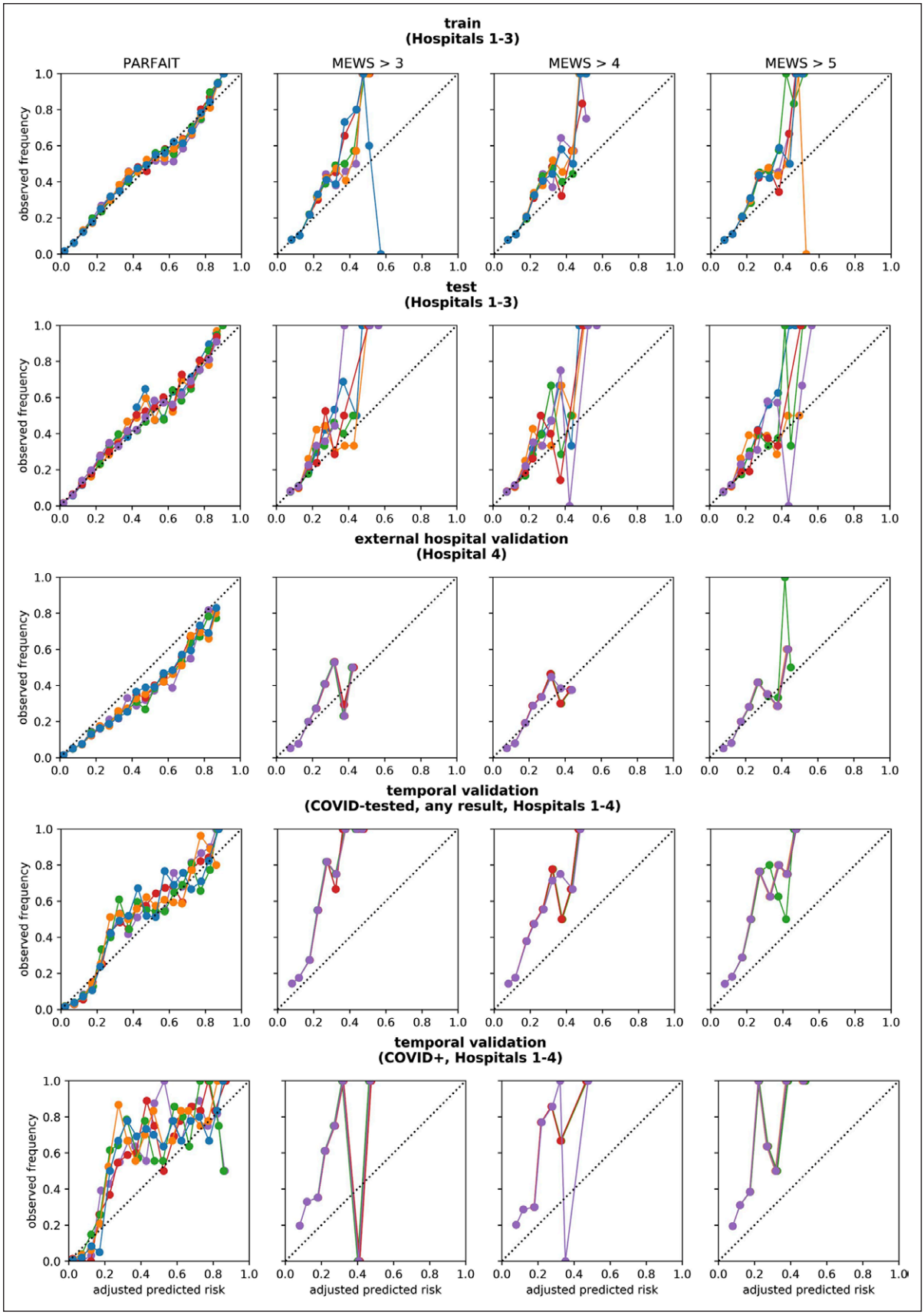

Figure 3. Calibration plots for each dataset, separated by method. COVID = coronavirus disease, MEWS = Modified Early Warning Score, PARFAIT = Prediction of Acute Respiratory Failure requiring advanced respiratory support in Advance of Interventions and Treatment.

is focused on ICU patients without a clear prediction of when the event will occur.

Most early warning scores predict ICU admission, cardiac arrest, or inhospital death—all of which are either late signs of dysfunction or have significant variation. Accurate Prediction of Prolonged Ventilation by Gong et al (25) predicts the need for IMV over 48 hours with a 48-hour event horizon. This provides a specific, more severe endpoint—3.08% in our dataset for IMV greater than 48 hours as compared with 8.4% for AdvRS. We chose a 3-hour event horizon as it provides an actionable time within a provider's shift (generally 12 hr) for care escalation, such as earlier diuresis or antibiotics. Since AdvRS also includes NIV and HHHF, it identifies patients earlier in their trajectory of decompensation and may mean an earlier intervention to further mitigate the risk of IMV. Furthermore, AdvRS offers more direct comparison of a patient's underlying oxygen requirements—provider behavior may mean that a patient may receive HHHF or NIV over IMV, so a pure endpoint of IMV would miss these patients.

PARFAIT performs substantially better in the COVID-19-tested cohort—and especially the COVID-19+ cohort, with an impressive AUROC of 0.93 and a sensitivity of 0.94 compared with MEWS greater than 3–5's AUROC 0.73–0.71 and a sensitivity of 0.29–0.09. Part of this performance increase is likely due to an increase in prevalence of AdvRS endpoints in the COVID-19 cohort. Furthermore, PARFAIT's false negative rate of 0.064—admittedly, in a cohort of 389 COVID-19+ patients—in conjunction with NPV 0.971 suggests that both positive and negative predictions may be clinically useful in stratifying patients—potentially permitting hospitals to anticipate resource need.

Despite the evaluated performance of PARFAIT, there are some limitations in this study. To avoid conflating this prediction with mortality, this study does not explicitly consider mortality as a predicted endpoint and misses ARF in a cardiopulmonary arrest that results in death before AdvRS is used. This model does capture patients with cardiopulmonary arrest that results in receiving AdvRS. We do not exclude patients on comfort care or do not resuscitate/do not intubate code status, which would alter clinician practice in intubating patients—but may result in patients going on NIV or HHHF, which would still be captured. As these data were cross-sectionally extracted from the EMR, the sampling bias represents the underlying population that is served by Emory Healthcare. Clinicians may order arterial blood gas (ABG)s in response to suspected decompensation, so ABG values can also be subject to availability bias. Additionally, given the exclusion of any patient encounters who received IMV in the OR, there is a lower surgical case representation, but this is an important exclusion since we want to predict unexpected respiratory decompensation. Finally, we are unable to robustly determine socioeconomic status, chronic medical conditions, and home oxygen requirements.

PARFAIT currently evaluates patients without knowledge of underlying comorbidities or home oxygen use. Further development to contextualize data by incorporating this knowledge could improve predictive accuracy. Model validation could be improved by further extra-system and prospective validation. Through temporal and external validations, we demonstrate PARFAIT's generalizability in providing robust predictions, even in unforeseen respiratory pandemics like COVID with changed practice patterns.

When implemented, PARFAIT generates predictions on an hourly basis, using votes of the past 8 hours. It is important to note that the current statistics are patient-level statistics, which may overestimate performance in the hourly predictions. With this work, PARFAIT is currently in its preliminary retrospective phase, requiring further real-time simulation and characterization prior to deployment. Next development steps involve creating real-time links to the EMR and silent evaluation to further characterize the performance of PARFAIT's hourly predictions. This evaluation should be carried out on both hourly metrics and patient level metrics to reflect on both the accuracy of identifying patients at risk of decompensation and the calibration of both event occurrence rate and time.

CONCLUSIONS

The PARFAIT risk stratification tool identifies patients at risk of developing ARF requiring AdvRS within a 3-hour event horizon. It performs better than MEWS with a higher PPV. PARFAIT demonstrates generalizability within the healthcare system both via external (external hospital, same healthcare system) and temporal validation, even in the face of an unpredicted

pandemic. Additional studies will be needed to determine whether automated prediction scores will result in improved clinical outcomes.

1 *Division of Pulmonary, Allergy, Critical Care, and Sleep Medicine, Department of Medicine, Emory University, Atlanta, GA.*

2 *Department of Biomedical Informatics, Emory University, Atlanta, GA.*

3 *Department of Biomedical Data Sciences, Leiden University Medical Centre, Leiden, The Netherlands.*

4 *Department of Information Technology and Digital Innovation, Leiden University Medical Centre, Leiden, The Netherlands.*

5 *Department of Biomedical Informatics, University of California San Diego, San Diego, CA.*

6 *Department of Surgery, Emory University, Atlanta, GA.*

7 *Department of Anesthesia, Emory University, Atlanta, GA.*

The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Dr. Wong performed experimental design, data processing, modeling, statistical analysis, article drafting, article revision, and final approval. Drs. Kamaleswaran, Holder, Buchman, and Blum were involved with experimental design, article revision, and final approval. Drs. Tabaie, Reyna, Josef, and Nemati were involved with data collection and article revision. Dr. Robichaux performed data collection. Ms. de Hond and Dr. Steyerberg were involved with statistical analysis and validation.

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).

Checklists: RECORD, TRIPOD.

Dr. Wong is supported by the National Institute of General Medical Sciences (NIGMS) 2T32GM095442 and the Clinical and Translational Science Award pilot informatics grant by National Center for Advancing Translational Sciences (NCATS) of the National Institutes of Health (NIH) under UL1TR002378. He holds equity and management roles in Ataia Medical. Dr. Kamaleswaran is supported by the Michael J. Fox Foundation (Grant No. 17267). Dr. Reyna is supported by NIH U54EB027690 and HHS0100201900015C. Dr. Josef is supported by the NIGMS 2T32GM095442. Dr. Holder is supported by the NIGMS under award number K23GM137182 for Advancing Translational Sciences of the NIH under Award Number UL1TR002378. Dr. Nemati is supported by the NIH (No. K01ES025445) and the Gordon and Betty Moore Foundation (No. GBMF9052). Dr. Buchman is supported by the Society of Critical Care Medicine and the Biomedical Advanced Research and Development Authority. He is an Editor in Chief for Critical Care Medicine and has recused himself from editorial influence on this article. Dr. Blum is supported by the NCATS of the NIH under Award Number UL1TR002378. He is a consultant for Clew Medical. The remaining authors have disclosed that they do not have any potential conflicts of interest.

For information regarding this article, E-mail: an-kwok.ian.wong@emory.edu

REFERENCES

1. Cartin-Ceba R, Kojicic M, Li G, et al: Epidemiology of critical care syndromes, organ failures, and life-support interventions in a suburban US community. *Chest* 2011; 140:1447–1455
2. Barrett ML, Smith MW, Elixhauser A, et al: Utilization of intensive care services, 2011: Statistical Brief# 185. *In*: Healthcare Cost and Utilization Project (HCUP) Statistical Briefs. 2006;15:1–4
3. Gedeon M: 52 - Pulmonary disorders. *In*: Parkland Trauma Handbook. Third Edition. Eastman AL, Rosenbaum DH, Thal ER (Eds). Philadelphia, PA, Mosby, 2009, pp 438–452
4. Churpek MM, Zadravecz FJ, Winslow C, et al: Incidence and prognostic value of the systemic inflammatory response syndrome and organ dysfunctions in ward patients. *Am J Respir Crit Care Med* 2015; 192:958–964
5. Chen J, Bellomo R, Flabouris A, et al: Delayed emergency team calls and associated hospital mortality: A multicenter study. *Crit Care Med* 2015; 43:2059–2065
6. Arentz M, Yim E, Klaff L, et al: Characteristics and outcomes of 21 critically ill patients with COVID-19 in Washington State. *JAMA* 2020; 323:1612–1614
7. Richardson S, Hirsch JS, Narasimhan M, et al: Presenting characteristics, comorbidities, and outcomes among 5700 patients hospitalized with COVID-19 in the New York City Area. *JAMA* 2020; 323:2052–2059
8. Bhatraju PK, Ghassemieh BJ, Nichols M, et al: Covid-19 in critically ill patients in the Seattle region - case series. *N Engl J Med* 2020; 382:2012–2022
9. Auld SC, Caridi-Scheible M, Blum JM, et al: ICU and ventilator mortality among critically ill adults with coronavirus disease 2019. *Crit Care Med* 2020; 48:e799–e804
10. Subbe CP, Kruger M, Rutherford P, et al: Validation of a modified early warning score in medical admissions. *QJM* 2001; 94:521–526
11. Green M, Lander H, Snyder A, et al: Comparison of the between the flags calling criteria to the MEWS, NEWS and the electronic Cardiac Arrest Risk Triage (eCART) score for the identification of deteriorating ward patients. *Resuscitation* 2018; 123:86–91
12. Kipnis P, Turk BJ, Wulf DA, et al: Development and validation of an electronic medical record-based alert score for detection of inpatient deterioration outside the ICU. *J Biomed Inform* 2016; 64:10–19
13. Chang HK, Wu CT, Liu JH, et al: Early detecting in-hospital cardiac arrest based on machine learning on imbalanced data. In 2019 IEEE International Conference on Healthcare Informatics (ICHI). IEEE, June 10, 2019, pp 1–10

14. Matam BR, Duncan H, Lowe D: Machine learning based framework to predict cardiac arrests in a paediatric intensive care unit: Prediction of cardiac arrests. *J Clin Monit Comput* 2019; 33:713–724
15. Mao Y, Chen Y, Hackmann G, et al: Medical data mining for early deterioration warning in general hospital wards. 2011 IEEE 11th International Conference on Data Mining Workshops. IEEE, December 11, 2011, pp 1042–1049
16. Admon AJ, Wunsch H, Iwashyna TJ, et al: Hospital contributions to variability in the use of ICUs among elderly medicare recipients. *Crit Care Med* 2017; 45:75–84
17. Dziadzko MA, Novotny PJ, Sloan J, et al: Multicenter derivation and validation of an early warning score for acute respiratory failure or death in the hospital. *Crit Care* 2018; 22:286
18. Leisman DE, Harhay MO, Lederer DJ, et al: Development and reporting of prediction models: Guidance for authors from editors of Respiratory, Sleep, and Critical Care Journals. *Crit Care Med* 2020; 48:623–633
19. Benchimol EI, Smeeth L, Guttmann A, et al; RECORD Working Committee: The reporting of studies conducted using observational routinely-collected health data (RECORD) statement. *PLoS Med* 2015; 12:e1001885
20. Moons KG, Altman DG, Reitsma JB, et al: Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): Explanation and elaboration. *Ann Intern Med* 2015; 162:W1–73
21. Chen T, Guestrin C: XGBoost: A scalable tree boosting system. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. San Francisco, CA, Association for Computing Machinery, August 13–17, 2016, pp 785–794
22. Wallace BC, Dahabreh IJ: Improving class probability estimates for imbalanced data. *Knowl Inf Syst* 2014; 41:33–52
23. Steyerberg EW, Vergouwe Y: Towards better clinical prediction models: Seven steps for development and an ABCD for validation. *Eur Heart J* 2014; 35:1925–1931
24. Steyerberg EW, Vickers AJ, Cook NR, et al: Assessing the performance of prediction models: A framework for traditional and novel measures. *Epidemiology* 2010; 21:128–138
25. Gong MN, Schenk L, Gajic O, et al: Early intervention of patients at risk for acute respiratory failure and prolonged mechanical ventilation with a checklist aimed at the prevention of organ failure: Protocol for a pragmatic stepped-wedged cluster trial of PROOFCheck. *BMJ Open* 2016; 6:e011347
26. Parreco J, Hidalgo A, Parks JJ, et al: Using artificial intelligence to predict prolonged mechanical ventilation and tracheostomy placement. *J Surg Res* 2018; 228:179–187
27. Martín-González F, González-Robledo J, Sánchez-Hernández F, et al: Success/failure prediction of noninvasive mechanical ventilation in intensive care units. Using multiclassifiers and feature selection methods. *Methods Inf Med* 2016; 55:234–241