# Risk assessment of acute respiratory failure requiring advanced respiratory support using machine learning

**Angier Allen**
Dascena (United States)
**Cecilia Zeng**
Dascena (United States)
**Chak Foon Tso**
Dascena (United States)
**Navan Singh**
Dascena (United States)
**Zohora Iqbal** ( ziqbal@dascena.com )
Dascena (United States)
**Misty M Attwood**
Dascena (United States)
**Veronica Gordon**
Dascena (United States)
**Cindy Wang**
Dascena (United States)
**Jana Hoffman**
Dascena (United States)

**Research Article**

**Keywords:** Acute respiratory failure, advanced respiratory support, machine learning, clinical decision support, artificial intelligence

**Posted Date:** June 8th, 2022

**DOI:** https://doi.org/10.21203/rs.3.rs-1668247/v1

## Abstract

# Background:

Acute respiratory failure (ARF) presents within a spectrum of clinical manifestations and illness severity, and mortality occurs in approximately 30% of patients who develop ARF. Early risk identification is imperative for implementation of prophylactic measures prior to ARF onset. In this study, we develop and validate a machine learning algorithm (MLA) to predict patients at risk of ARF requiring advanced respiratory support.

# Methods:

This retrospective study used data from 155,725 patient electronic health records obtained from five United States community hospitals. An XGBoost classifier was developed using patient EHR data to produce risk scores at 3-hour intervals to predict the risk of ARF within 24 hours. An alert was generated only once prior to ARF onset, defined by implementation of advanced respiratory support, for patients whose risk score exceeded a predefined threshold. We used a novel time-sensitive area under the receiver operating characteristic (tAUROC) curve that integrated the timing of the alert relative to ARF onset to evaluate the accuracy of the MLA. The MLA was assessed on two testing sets and compared with oxygen saturation ($SpO_2$) measurement and the modified early warning score (MEWS).

# Results:

The MLA demonstrated significantly higher eSensitivity and specificity operating points on the temporal testing and external validation sets (tAUROC of 0.858/0.883, respectively) than $SpO_2$ (0.771/0.810) and MEWS (0.676/0.774) for prediction of ARF requiring advanced respiratory support. The MLA also achieved lower false positive rates than $SpO_2$ and MEWS at these operating points.

# Conclusions:

The MLA predicts patients at risk of ARF requiring advanced respiratory support and achieves higher accuracy and produces earlier alerts than the use of $SpO_2$ or MEWS. Importantly for clinical practice, the MLA has a lower false positive rate than these comparators while maintaining high sensitivity and specificity.

## Background

Acute respiratory failure (ARF) is broadly characterized as inadequate gas exchange, where the respiratory system is unable to meet the oxygenation, ventilation, or metabolic demands of the body (1, 2). The leading causes of ARF in adults include pneumonia, congestive heart failure, chronic obstructive pulmonary disease, acute respiratory distress syndrome, sepsis, asthma, drug ingestion, and trauma (3). As the most common cause of admission to the intensive care unit (ICU) (4), ARF affects more than 50% of ICU patients (5). Severe ARF that requires mechanical ventilation has mortality rates of 34–37% (6, 7). In the United States, the number of hospitalizations due to ARF increased from ~ 1 million to nearly 2 million from 2001 to 2009 (3). The incidence of ARF continues to increase, possibly due to an aging population along with septicemia and seasonal fluxes of influenza and pneumonia (8). Signs of respiratory deterioration may occur before ARF onset, and early recognition and application of prophylactic interventions may improve outcomes and reduce mortality (9). Hence, there is substantial need for accurate ARF risk assessment.

Due to the high morbidity and mortality associated with ARF, early identification and intervention is beneficial. To this end, several studies have proposed improving and/or reducing the time to ARF diagnosis by applying techniques including lung ultrasounds (10, 11), CT-scans (11), and chest x-rays with electronic health records (EHR) data (12). However, these tests are

typically performed after suspicion of ARF, and studies have demonstrated that delayed or inappropriate identification may lead to increased mortality (13, 14). In the absence of ARF risk prediction tools, identification of ARF relies upon methods that are costly (e.g., imaging) (11), are not considered primary diagnostic measures (11), or that require physician assessment in combination with clinical variables (15). Several ARF risk assessment tools have been validated for use in post-surgical populations to assess the risk of respiratory complications (16–18). However, there is currently a need for physician-unassisted monitoring and identification of patients at risk of respiratory decompensation in all hospital settings given current staffing challenges, particularly with the dwindling number of experienced nursing staff (19).

In this study, we developed a continually-running machine learning algorithm (MLA) using a minimal number of features that are readily available in the EHR to assess the risk of ARF in hospitalized patients requiring advanced respiratory support. This clinical decision support tool could enable clinicians to implement earlier interventions or preventative measures for patients at risk of ARF.

## Methods

# Data collection

Demographics, vitals, laboratory test results, and comorbidities data were extracted for adult patients from the EHRs of five United States community hospitals, including a large academic research hospital. Patient data was de-identified in compliance with the Health Insurance Portability and Accountability Act. As such, this research does not constitute human subjects research as per the definition put forth in 45 Code of Federal Regulations 46 and did not require Institutional Review Board approval (20).

# Gold standard

Management of ARF varies with severity, ranging from supplemental $O_2$ therapy using nasal cannulas, high-flow nasal cannulas, venturi masks, and non-rebreather reservoir masks, to non-invasive ventilation (NIV) using positive pressure ventilation with continuous positive airway pressure (CPAP) and bilevel positive airway pressure (BiPAP) (13, 21–24). When non-invasive techniques are not sufficient, patients are intubated and provided mechanical ventilation (25, 26). Mechanical ventilation and intubation carry risks including infections and physical injuries (27). In case of prolonged ARF, a tracheostomy may also be considered (28).

ARF patients requiring advanced respiratory support were identified by specific ventilation procedures: ECMO, tracheostomy, mechanical ventilation, high-flow nasal cannulas, CPAP, BiPAP, and non-rebreather masks. ARF onset time was determined by the start of these procedures. The list of string searches for oxygen therapy in patient EHR are listed in **Additional File 1**. For added verification in determining ventilation methods from the patient EHR, we also required a minimum saturated oxygen level ($SpO_2$) of less than 96% from − 12 to + 6 hours around the ventilation time (29). Non-invasive ventilation support, such as low or moderate flow nasal cannula (1–6 L and 6–15 L oxygen flow, respectively), was not considered advanced respiratory support. The use of these advanced respiratory support measures to identify ARF patients were determined through expert clinician discussions (see **Additional File 2** for details) and have been employed in previous clinical studies (4, 30). Advanced respiratory support devices that provide higher levels of oxygen support can be used as a surrogate for ARF severity (4), and provide a more direct endpoint of respiratory decompensation than invasive mechanical ventilation or admission to the ICU (30). Therefore, this study defines the positive class ARF encounters as all patient visits where the patient used advanced respiratory support, and the negative class, non-ARF encounters, as all patient visits where the patient did not use advanced respiratory support.

# Data processing

The dataset was temporally divided into a training and a testing set from four hospital sites (sites A-D), and a fifth site (E) was used for external validation testing. The training dataset included data from sites A-D between April 1, 2020 and November 30, 2021. The temporal testing set included data obtained from sites A-C between December 1, 2021 and February 20, 2022.

External validation was conducted using data obtained from the fifth, independent site (E) between April 1, 2020 and February 20, 2022. Data from the training, testing and external validation sets did not overlap. All encounters of patients aged 18 to 100 years that had at least one measurement of each required feature (described below) present before ARF onset were included in the study. Patients that had a pre-existing ARF condition were excluded from the study.

# Model inputs

The model required minimal features, including demographic features (age and sex), vital signs (systolic and diastolic blood pressure, heart rate, respiratory rate, and temperature), $SpO_2$, and COVID-19 status as a boolean feature. Optional features were included in model inputs as available (Table 1). Information on data missingness can be found in **Additional File 3.**

Table 1
Required and optional model input features. COVID-19 status was a boolean feature.

| Required features | Optional features |
|---|---|
| ● Age | ● Hematocrit |
| ● Systolic blood pressure | ● Hemoglobin |
| ● Diastolic blood pressure | ● White blood cell count |
| ● Heart rate | ● Red blood cell count |
| ● Respiratory rate | ● Platelet count |
| ● Temperature | ● Creatinine |
| ● Oxygen saturation ($SpO_2$) | ● Blood urea nitrogen (BUN) |
| ● COVID-19 status | ● Blood pH |
| | ● Partial pressure of oxygen ($PaO_2$) |

# Machine learning model

# XGB model

The XGBoost classifier (31) was selected to create the model for analyzing inputs and assessing risk of ARF. XGBoost implements gradient boosting, which is an ensemble learning technique that combines multiple decision trees to produce respiratory decompensation risk scores. At each iteration of the algorithm, regression trees are sequentially combined to improve on errors of previous iterations. The XGBoost classifier was chosen due to its reliability and scalability. It is able to process data with missing values without first performing imputation (31, 32). XGBoost has a number of parameters that can be tuned to control class imbalance and also avoid overfitting. To avoid overfitting to a population, we used $\max_{d} epth = 4$ and $\min_{\chi} ld_{w}eight = 50$. All other parameters were left in their default configuration.

# Algorithm design and training

Alert timing is an important aspect of designing clinical decision support systems. Discussions with clinical experts identified alerts prior to ARF onset, and particularly within a 24-hour period before onset, to be the most useful alerting window (**Additional File 2**). This would provide sufficient notice to implement treatment and prepare patients for potential ICU transfer, care escalation, or mechanical ventilation. The MLA was deployed at $T_0$, when at least one measurement for each required model input was present. Risk scores predicting ARF onset within the next 24 hours were generated at 3-hour intervals with updated EHR data. If a feature was not updated in that 3-hour interval, then exact match forward imputation was used for that measurement. This process was repeated from $T_0$ to ARF onset for ARF-positive patients, and $T_0$ to time of discharge for ARF-negative patients. If a 3-hour interval fell within 24 hours of ARF onset, the model was trained to predict positive for ARF.

Otherwise, it was trained to predict negative for ARF. The training strategy is illustrated in Fig. 1. Although the MLA generated multiple risk scores during a patient's length of stay, it only produced an alert at the first instance the risk score crossed the predefined alert threshold, which was determined by optimized sensitivity and specificity values in the training set.

# Model performance evaluation

Traditional methods of evaluating algorithm performance using area under the receiver operator characteristic (AUROC) curve measure the accuracy of a model (33–35), but do not include analysis on when an alert is generated, i.e., the timing of the alert (36, 37). This is relevant as alerts generated before the onset of a disease may provide greater opportunity for clinicians to apply prophylactic measures to patients.

To incorporate the concept of alert timing in the evaluation of algorithm performance, we introduced modified definitions of true positive and false negative cases. A true positive case was defined as an ARF encounter where an alert was generated prior to ARF onset, thus only *early* cases were included. A false negative case was defined as an ARF encounter where an alert was not provided before the onset of ARF. Note that late alerts that were generated *after* the onset of ARF were considered false negatives. Evaluation metrics, including sensitivity (eSensitivity) were adjusted based on these modified definitions (Table 2). eSensitivity was defined as the proportion of ARF encounters for which the algorithm provided an alert before ARF onset out of the total number of ARF encounters.

# Accounting for multiple scores in receiver operating characteristic curves

As the model produced multiple scores for each encounter, we first converted these into a final prediction before using them to build a receiver operating characteristic (ROC) curve. A traditional ROC curve is constructed by varying the prediction threshold to produce all possible operating points (sensitivity and specificity) for the model. We mirrored this computation by varying the alert threshold to produce a range of operating points. For each alert threshold, the alert time for an encounter was given as the first time a score surpassed the selected alert threshold. This alert time was compared with ARF onset time, and the encounter was categorized in one of the four classifications (early true positive, false positive, etc.) described above and in Table 2, which was used to compute eSensitivity and specificity. Computing eSensitivity and specificity for all possible alert thresholds yields the time-sensitive ROC (tROC) curves in Fig. 3, and the area under the tROC curve (tAUROC) is reported in Table 4. tAUROC represents the time-sensitive accuracy of the model by evaluating area under the receiver operating curve of eSensitivity to (1-specificity).

Unlike traditional ROC curves produced from a single prediction, one drawback of this approach is that for a model producing random scores at each point in time, the tROC curve will not be a diagonal line and the tAUROC will not be 0.5. This makes interpretation of tAUROC more complex than the standard case. To better contextualize the models' performance, expected baseline tROC curves and their associated tAUROC values are provided in Fig. 3.

# Evaluating the timing of alerts

The timing of alerts was evaluated to determine the number of alerts that were generated within clinically relevant windows to implement prophylactic measures. We investigated alerts produced during two different windows: an *early* alert was one generated at any time prior to the onset of ARF, and a *timely* alert was one that was produced within the 24 hours prior to ARF onset. Clinical feedback suggested that timely alerts are the most clinically optimizable for preventive measures (**Additional File 2**). We then defined *timeliness* as the proportion of early true positive alerts generated during this ideal window.

Table 2
Definitions of performance metrics

| Definitions incorporating timeliness of alerts |
|---|
| **eTP**: early true positive; ARF encounters alerted before ARF onset |
| **FP**: false positive; non-ARF encounters that received an alert |
| **TN**: true negative; non-ARF encounters that did not receive an alert |
| **eFN**: early false negative; ARF encounters that did not receive an alert, or received an alert after ARF onset |
| **eSensitivity**: early sensitivity; # of ARF encounters correctly alerted before ARF onset/ # of total ARF encounters); (eTP/(eTP + eFN)) |
| **Specificity**: # of non-ARF patients that was not alerted/ # of total negative patients (TN/(FP + TN)) |
| **tAUROC**: time-sensitive area under the receiver operating curve; measure of correctness; plots eSensitivity to (1-specificity) |
| **Early alert**: alert generated at any time prior to the onset of ARF |
| **Timely alert**: alert was one that was produced within the 24 hours prior to ARF onset |
| **Timeliness**: the proportion of early true positive (eTP) alerts generated within 24 hours prior to ARF onset |

# Comparators

To establish a baseline performance for the model, we compared the MLA with two different comparators: patient $SpO_2$ measurements below a specific threshold, and the Modified Early Warning Score (MEWS) (38). Monitoring $SpO_2$ is widely used in clinical practice for patient assessment (39) and evidence suggests that the $SpO_2$ target range for healthy, non-smoking adults should be greater than 94% (40). Guidelines advise oxygen should be prescribed to achieve saturation levels greater than 94% (41). In addition to literature support, clinician discussions identified $SpO_2$ levels as one of the major determinants in assessing patient status (**Additional File 2**). We defined the threshold for the $SpO_2$ score $\leq$ 93%, where lower $SpO_2$ values indicate higher risk.

MEWS is a commonly used scoring system that identifies likely patient deterioration and mortality and has been used in previous machine learning studies with input adjustments (42, 43). It is typically computed from patient vital signs and includes a subcomponent, the AVPU (alert, verbal, pain, unresponsive) assessment, that was not readily available in the dataset and thus not included in the MEWS calculation. The adjusted MEWS was calculated from minimal features including heart rate, temperature, respiratory rate and systolic blood pressure **(Additional File 4)**. The MEWS score ranged from 0 to 11, where an elevated score indicated an increased risk for clinical instability. We chose a threshold of $\geq$ 2 for comparison.

# SHAP analysis

Model feature importance was evaluated using SHAP (SHapley Additive exPlanations) plots (44). SHAP plots analyze feature correlations and distributions at a global ranking level and aid in interpretability of the features and algorithm model. The plots present the predictive value of each feature input and its positive or negative correlation with predicting ARF patients needing advanced respiratory support.

## Results

# Demographics information

Of the 175,576 hospitalized patient encounters in the datasets, 155,725 encounters met the inclusion criteria and were included in the training and testing sets (Fig. 2). The training set consisted of 4,456 ARF encounters and 111,074 non-ARF encounters to achieve a prevalence of 3.9%. The temporal testing set included 343 ARF encounters and the prevalence was

4.2%. The external validation set included 842 ARF encounters and the prevalence was 2.6%. The demographics information of patient encounters included in the analysis is presented in Table 3.

Table 3
Patient encounter (enc) characteristics. ARF, acute respiratory failure; ED, emergency department; ICU, intensive care unit. Unreported race is categorized as "Unknown race."

| | | Training Set | | | Temporal Testing Set | | | External Validation Set | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Feature | Demo-graphics | Non-ARF enc (n = 111,074) | ARF enc (n = 4,456) | p-value | Non-ARF enc (n = 7,820) | ARF enc (n = 343) | p-value | Non-ARF enc (n = 31,190) | ARF enc (n = 842) | p-value |
| Age | 18–39 | 35,748 (32.2%) | 258 (5.8%) | < 0.001 | 1,943 (24.8%) | 24 (7.0%) | < 0.001 | 8,261 (26.5%) | 44 (5.2%) | < 0.001 |
| | 40–59 | 30,828 (27.8%) | 1,075 (24.1%) | < 0.001 | 2,062 (26.4%) | 54 (15.7%) | < 0.001 | 8,748 (28.0%) | 145 (17.2%) | < 0.001 |
| | 60–79 | 33,230 (29.9%) | 2,349 (52.7%) | < 0.001 | 2,726 (34.9%) | 183 (53.4%) | < 0.001 | 9,205 (29.5%) | 375 (44.5%) | < 0.001 |
| | 80+ | 11,268 (10.1%) | 774 (17.4%) | < 0.001 | 1,089 (13.9%) | 82 (23.9%) | < 0.001 | 4,976 (16.0%) | 278 (33.0%) | < 0.001 |
| Sex | Male | 48,946 (44.1%) | 2,373 (53.3%) | < 0.001 | 3,711 (47.5%) | 200 (58.3%) | 0.024 | 13,608 (43.6%) | 456 (54.2%) | < 0.001 |
| | Female | 62,128 (55.9%) | 2,083 (46.7%) | < 0.001 | 4,109 (52.5%) | 143 (41.7%) | 0.022 | 17,582 (56.4%) | 386 (45.8%) | < 0.001 |
| Ethnicity/Race | White | 25,677 (23.1%) | 998 (22.4%) | 0.386 | 3,132 (40.1%) | 95 (27.7%) | 0.001 | 3,874 (12.4%) | 287 (34.1%) | < 0.001 |
| | Black | 2,054 (1.8%) | 61 (1.4%) | 0.019 | 317 (4.1%) | 6 (1.7%) | 0.032 | 618 (2.0%) | 40 (4.8%) | < 0.001 |
| | Asian | 98 (0.1%) | 1 (0.0%) | 0.190 | 13 (0.2%) | 0 (0.0%) | 1.000 | 126 (0.4%) | 10 (1.2%) | 0.003 |
| | Other Race | 731 (0.7%) | 19 (0.4%) | 0.069 | 99 (1.3%) | 0 (0.0%) | 0.036 | 1,346 (4.3%) | 81 (9.6%) | < 0.001 |
| | Unknown Race | 82,514 (74.3%) | 3,377 (75.8%) | 0.395 | 4,259 (54.5%) | 242 (70.6%) | 0.003 | 25,226 (80.9%) | 424 (50.4%) | < 0.001 |
| | Hispanic | 2,066 (1.9%) | 36 (0.8%) | < 0.001 | 330 (4.2%) | 1 (0.3%) | < 0.001 | 1,284 (4.1%) | 73 (8.7%) | < 0.001 |
| | Non-hispanic | 22,738 (20.5%) | 665 (14.9%) | < 0.001 | 3,232 (41.3%) | 100 (29.2%) | 0.002 | 4,519 (14.5%) | 344 (40.9%) | < 0.001 |
| | Unknown Ethnicity | 86,270 (77.7%) | 3,755 (84.3%) | < 0.001 | 4,258 (54.5%) | 242 (70.6%) | 0.003 | 25,387 (81.4%) | 425 (50.5%) | < 0.001 |
| Hospital ward | in ICU | 2,938 (2.6%) | 634 (14.2%) | < 0.001 | 167 (2.1%) | 38 (11.1%) | < 0.001 | 819 (2.6%) | 106 (12.6%) | < 0.001 |
| | in ED | 40,895 (36.8%) | 1,939 (43.5%) | < 0.001 | 3,110 (39.8%) | 99 (28.9%) | 0.005 | 12,661 (40.6%) | 464 (55.1%) | < 0.001 |
| COVID-19 status | COVID19 | 2,586 (2.3%) | 620 (13.9%) | < 0.001 | 365 (4.7%) | 65 (19.0%) | < 0.001 | 1,621 (5.2%) | 206 (24.5%) | < 0.001 |

# Performance results

The performance of the MLA was evaluated in comparison to the two comparators, $SpO_2$ and MEWS, on the temporal testing set and external validation set (Fig. 3 and Table 4). Figure 3 presents the tROCs, including the baseline random model, which deviates from the diagonal and does not have an area under the curve of 0.5 for an alert time-insensitive ROC. The MLA achieved tAUROC values of 0.858 and 0.883 on the temporal and external sets, respectively, in comparison with $SpO_2$ tAUROC values of 0.771 and 0.810; MEWS achieved 0.676 and 0.774.

Table 4
Performance of the machine learning algorithm (MLA) in comparison to $SpO_2$ and modified early warning score (MEWS) comparators on the temporal testing set and external validation set. The modified performance metrics included the time-sensitive area under the receiver operating characteristic (tAUROC) and early sensitivity (eSensitivity) values. The threshold for $SpO_2$ was ≤ 93% and the threshold for the MEWS comparator was ≥ 2.0.

| | Temporal testing set | | | External validation set | | |
|---|---|---|---|---|---|---|
| **Method** | **tAUROC** | **eSensitivity** | **Specificity** | **tAUROC** | **eSensitivity** | **Specificity** |
| **MLA** | 0.858 (0.844–0.872) | 0.776 | 0.780 | 0.883 (0.871–0.895) | 0.748 | 0.864 |
| **$SpO_2$** | 0.771 (0.746–0.796) | 0.668 | 0.733 | 0.810 (0.794–0.824) | 0.582 | 0.866 |
| **MEWS** | 0.676 (0.650–0.696) | 0.653 | 0.660 | 0.774 (0.756–0.789) | 0.818 | 0.641 |

# SHAP analysis

As demonstrated in Fig. 4, the SHAP analysis of feature importance showed age, $SpO_2$, blood urea nitrogen (BUN), respiratory rate, heart rate, and pH were of highest value to model prediction.

# Timing of alerts

Of the 842 ARF encounters in the external validation set, 630 were alerted prior to the onset of ARF (**Additional File 5).** At a sensitivity of 74.8%, the algorithm achieved a false positive rate (FPR) of 13.6%; $SpO_2$ at a reduced sensitivity of 58.2% had an FPR of 13.4%, while the MEWS at increased sensitivity of 81.8% achieved an FPR of 35.9%. A similar trend is seen for the temporal testing set.

The 630 *early* alerts the MLA produced on the external validation set had a median alerting time of 23.0 hours before ARF onset (Fig. 5). There were a total of 212 false negatives in the external validation set, indicating a false negative rate of 25.2%. For additional analysis of alert timing, we calculated that, of the 212 false negatives, 105 ARF encounters would have received an alert after the onset of ARF. The 266 early alerts generated on the temporal testing set had a median alerting time of 21.0 hours prior to ARF onset (Fig. 5). There were 77 false negatives, of which 31 would have received an alert after ARF onset, indicating a false negative rate of 22.4%.

## Discussion

In this retrospective study, we developed and validated an algorithm that identifies patients at risk of ARF requiring advanced respiratory support. The algorithm continually produced risk scores at 3-hour intervals based on updated patient EHR data to predict the risk of ARF within the next 24 hours. An alert was generated only once for patients when the risk score first exceeded a predefined threshold prior to ARF onset. The MLA demonstrated superior performance at higher sensitivity and specificity in comparison to $SpO_2$ and MEWS comparators when using time-sensitive ROC evaluation methods. Importantly, we demonstrate a greater number of *early* alerts before ARF onset with a lower false positive rate, reducing false alarms.

An easily implementable, automatic clinical decision support tool that does not require physician input or interrupt clinical workflow may supplement clinical assessment in hospitals lacking adequate staffing or resources. The continually running

algorithm design and analysis of new patient data every three hours to predict ARF risk within a 24-hour prediction window would provide clinically relevant alerts to health care providers. MLA defined ARF advanced respiratory support treatments include NIV procedures and high flow nasal cannula, allowing identification of patients early in their trajectory of decompensation, and enabling prophylactic measures. This also would permit earlier consideration of appropriate ventilation escalation, including techniques that are protective against lung injury (45–48).

The MLA performed with higher sensitivity and specificity on the temporal testing and external validation sets (tAUROC of 0.858/0.883, respectively) than $SpO_2$ (0.771/0.810) and the MEWS score (0.676/0.774). Other studies have created algorithms predicting ARF onset (30, 49); however, direct comparisons are difficult due to different methodologies and evaluation metrics. Wong et al. (30) developed an XGBoost ML model with a similar definition of ARF that provided a single alert with a 3 hour prediction window from the majority vote of 8 predictions over the span of 8 hours. In comparison, we generated risk scores with a 24-hour prediction window at 3-hour intervals from updated EHR data until a risk score exceeded a specific threshold and an alert was generated. Although the algorithm from Wong et al. outperformed MEWS with an AUROC of 0.85 versus 0.57, it is a complex algorithm with 70 input variables, many of which may not be immediately available in many healthcare systems, limiting its applicability. Kim et al. (49) designed an MLA where acute respiratory failure was defined as endotracheal intubation to make risk predictions within a window of 6 hours to 1 hour prior to onset using nine features. While they also achieved high AUROC values, their definition of ARF included only mechanical intubation, and a shorter prediction window.

The most important features that contributed to the model's performance were age, $SpO_2$, BUN, pH, and respiratory rate for both the temporal testing and external validation sets. Age is known to contribute to ARF due to age-related structural changes of the respiratory system (13, 50). BUN is a marker of renal function, (51) and its importance may be correlated with the increasing body of evidence that suggests lung-kidney interactions are involved in renal consequences of ARF (52). $SpO_2$ has been previously identified in literature (39, 40) and in discussions with clinicians as critical for monitoring patient respiratory status. $PaO_2$ was of lower importance in MLA prediction outcomes in both data sets. One reason may be that $SpO_2$ is a non-invasive tool and typically monitored continually with frequent EHR updates. In contrast, $PaO_2$ is a laboratory test with a high degree of data missingness, that may be determined only after the clinician already suspects a decline in patient respiratory status. While pH had high data missingness similar to $PaO_2$, the importance of pH appears to be more informative for the model than $PaO_2$. Any measurement of pH resulted in a positive SHAP value for predictions, while a missing value resulted in a slightly negative value. Covid-19 positive status was also of high importance in the MLA prediction of ARF requiring respiratory support, as may be expected (53–55).

An important aspect of designing clinical decision support systems is the clinical relevancy and the timing of alerts. From literature and discussions with clinician experts, *early* alerts prior to ARF onset (30), and especially within 24 hour before onset, are considered to be most useful (**Additional File 2**). With *early* alerts, health care providers may implement prophylactic or preventive therapies to mitigate ARF onset. Therefore, we evaluated the algorithm performance with respect to the time of ARF onset. The MLA achieved a similar or lower false positive rate than the comparators. While $SpO_2$ achieved a similar false positive rate (13.4%) to the MLA (13.6%), the number of true positive cases identified before ARF onset by the MLA was much higher, and the MEWS comparator produced over 2.5-times the number of false positives than the algorithm.

There are several strengths to this analysis. The patient data used in this study was obtained from multiple sites from throughout the United States and represent a heterogeneous patient population. The analysis evaluates the model on both a temporally distinct set and an external validation set, demonstrating generalizability. This study is limited by its retrospective nature, and the lack of prospective validation to ascertain clinician response to alerts in practice. We cannot predict how the MLA may perform in different populations or settings. The MEWS score was calculated without AVPU assessment, as it was not available in the dataset. Therefore, the performance of the MEWS score may be underestimated in this study.

# Conclusions

In this analysis, we present a novel algorithm to identify patients at risk of ARF requiring advanced respiratory support using only EHR data. Integrating continual updates from new patient data and generating risk scores every three hours, the MLA accurately predicted ARF onset within the next 24 hours. This high-performing classifier produces more accurate and clinically relevant alerts for patients at risk of ARF than $SpO_2$ measurements or MEWS. Importantly for relevance and utility to health care providers, high sensitivity and specificity were achieved for MLA alerts prior to ARF onset. This may enable early intervention to improve patient outcomes.

## Abbreviations

**ABG**
arterial blood gas
**ARF**
acute respiratory failure
**BiPAP**
bilevel positive airway pressure
**CPAP**
continuous positive airway pressure
**EHR**
electronic health record
**ECMO**
extracorporeal membrane oxygenation
**eSensitivity**
early sensitivity; ratio of the number of encounters correctly alerted before acute respiratory failure onset to the number of total ARF encounters
**FPR**
false positive rate
**ICU**
intensive care unit
**MEWS**
modified early warning score
**MLA**
machine learning algorithm
**NIV**
non-invasive ventilation
**$PaO_2$**
partial pressure of arterial oxygen
**ROC**
receiver operating characteristic
**SHAP**
SHapley Additive exPlanations
**$SpO_2$**
saturated oxygen level
**tAUROC**
time-sensitive area under the receiver operating characteristic

## Declarations

# Ethics approval and consent to participate

Patient data was de-identified in compliance with the Health Insurance Portability and Accountability Act. As such, this research does not constitute human subjects research as per the definition put forth in 45 Code of Federal Regulations 46 and did not require Institutional Review Board approval.

# Consent for publication

Not applicable

# Availability of data and materials

The data that support the findings of this study have restrictions that apply to the availability of these data and are not publicly available. The MLA code developed in this study is proprietary and not publicly available.

# Competing interests

All authors who have affiliations listed with Dascena (Houston, Texas, U.S.A) are employees or contractors of Dascena.

# Funding

Not applicable

# Author’s contributions

AA, CFT, CZ, CW, VG contributed to the conception and design of the work; NS contributed to the acquisition and analysis of data; MMA, ZI, AA, CZ, and JH contributed to the writing and revising of the work.

# Acknowledgements

We would like to thank the panel of medical experts and practitioners that participated in discussions and provided valuable insight and feedback. The authors thank Gina Barnes for her research support.

## References

1. Canet J, Gallart L. Postoperative respiratory failure: pathogenesis, prediction, and prevention. Curr Opin Crit Care. 2014 Feb;20(1):56–62.
2. Burt CC, Arrowsmith JE. Respiratory failure. Surg Oxf. 2009 Nov 1;27(11):475–9.
3. Stefan MS, Shieh MS, Pekow PS, Rothberg MB, Steingrub JS, Lagu T, et al. Epidemiology and outcomes of acute respiratory failure in the United States, 2001 to 2009: A national survey. J Hosp Med. 2013;8(2):76–82.
4. Wong AKI, Cheung PC, Kamaleswaran R, Martin GS, Holder AL. Machine Learning Methods to Predict Acute Respiratory Failure and Acute Respiratory Distress Syndrome. Front Big Data [Internet]. 2020 [cited 2022 Mar 2];3. Available from: https://www.frontiersin.org/article/10.3389/fdata.2020.579774
5. Vincent JL, Akça S, de Mendonça A, Haji-Michael P, Sprung C, Moreno R, et al. The Epidemiology of Acute Respiratory Failure in Critically Ill Patients. Chest. 2002 May 1;121(5):1602–9.
6. Needham DM, Bronskill SE, Sibbald WJ, Pronovost PJ, Laupacis A. Mechanical ventilation in Ontario, 1992–2000: Incidence, survival, and hospital bed utilization of noncardiac surgery adult patients*. Crit Care Med. 2004

Jul;32(7):1504–9.

7. Carson SS, Cox CE, Holmes GM, Howard A, Carey TS. The Changing Epidemiology of Mechanical Ventilation: A Population-Based Study. J Intensive Care Med. 2006 May 1;21(3):173–82.
8. Parcha V, Kalra R, Bhatt SP, Berra L, Arora G, Arora P. Trends and Geographic Variation in Acute Respiratory Failure and ARDS Mortality in the United States. CHEST. 2021 Apr 1;159(4):1460–72.
9. Gong MN, Schenk L, Gajic O, Mirhaji P, Sloan J, Dong Y, 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 Jun 10;6(6):e011347.
10. Lichtenstein DA, Mezière GA. Relevance of Lung Ultrasound in the Diagnosis of Acute Respiratory Failure*: The BLUE Protocol. Chest. 2008 Jul 1;134(1):117–25.
11. Wallbridge P, Steinfort D, Tay TR, Irving L, Hew M. Diagnostic chest ultrasound for acute respiratory failure. Respir Med. 2018 Aug 1;141:26–36.
12. Jabbour S, Fouhey D, Kazerooni E, Wiens J, Sjoding MW. Combining chest X-rays and electronic health record (EHR) data using machine learning to diagnose acute respiratory failure. J Am Med Inform Assoc. 2022 Mar 10;ocac030.
13. Delerme S, Ray P. Acute respiratory failure in the elderly: diagnosis and prognosis. Age Ageing. 2008 May 1;37(3):251–7.
14. Boniatti MM, Azzolini N, Viana MV, Ribeiro BSP, Coelho RS, Castilho RK, et al. Delayed Medical Emergency Team Calls and Associated Outcomes*. Crit Care Med. 2014 Jan;42(1):26–30.
15. Shebl E, Burns B. Respiratory Failure. In: StatPearls [Internet]. Treasure Island (FL): StatPearls Publishing; 2022 [cited 2022 Apr 18]. Available from: http://www.ncbi.nlm.nih.gov/books/NBK526127/
16. Mazo V, Sabaté S, Canet J, Gallart L, de Abreu MG, Belda J, et al. Prospective External Validation of a Predictive Score for Postoperative Pulmonary Complications. Anesthesiology. 2014 Aug 1;121(2):219–31.
17. Gupta H, Gupta PK, Fang X, Miller WJ, Cemaj S, Forse RA, et al. Development and Validation of a Risk Calculator Predicting Postoperative Respiratory Failure. Chest. 2011 Nov 1;140(5):1207–15.
18. Canet J, Gallart L, Gomar C, Paluzie G, Vallès J, Castillo J, et al. Prediction of Postoperative Pulmonary Complications in a Population-based Surgical Cohort. Anesthesiology. 2010 Dec 1;113(6):1338–50.
19. Griffiths P, Recio-Saucedo A, Dall'Ora C, Briggs J, Maruotti A, Meredith P, et al. The association between nurse staffing and omissions in nursing care: A systematic review. J Adv Nurs. 2018 Jul;74(7):1474–87.
20. Protections (OHRP) O for HR. Exemptions (2018 Requirements) [Internet]. HHS.gov. 2021 [cited 2022 Jan 21]. Available from: https://www.hhs.gov/ohrp/regulations-and-policy/regulations/45-cfr-46/common-rule-subpart-a-46104/index.html
21. Hill NS, Brennan J, Garpestad E, Nava S. Noninvasive ventilation in acute respiratory failure. Crit Care Med. 2007 Oct;35(10):2402–7.
22. Tomii K, Seo R, Tachikawa R, Harada Y, Murase K, Kaji R, et al. Impact of noninvasive ventilation (NIV) trial for various types of acute respiratory failure in the emergency department; decreased mortality and use of the ICU. Respir Med. 2009 Jan 1;103(1):67–73.
23. Grieco DL, Maggiore SM, Roca O, Spinelli E, Patel BK, Thille AW, et al. Non-invasive ventilatory support and high-flow nasal oxygen as first-line treatment of acute hypoxemic respiratory failure and ARDS. Intensive Care Med. 2021 Aug 1;47(8):851–66.
24. Qaseem A, Etxeandia-Ikobaltzeta I, Fitterman N, Williams JW, Kansagara D. Appropriate Use of High-Flow Nasal Oxygen in Hospitalized Patients for Initial or Postextubation Management of Acute Respiratory Failure: A Clinical Guideline From the American College of Physicians. Ann Intern Med. 2021 Jul 20;174(7):977–84.
25. Fan E, Del Sorbo L, Goligher EC, Hodgson CL, Munshi L, Walkey AJ, et al. An Official American Thoracic Society/European Society of Intensive Care Medicine/Society of Critical Care Medicine Clinical Practice Guideline: Mechanical Ventilation in Adult Patients with Acute Respiratory Distress Syndrome. Am J Respir Crit Care Med. 2017;195(9):1253–63.
26. Azoulay É, Thiéry G, Chevret S, Moreau D, Darmon M, Bergeron A, et al. The Prognosis of Acute Respiratory Failure in Critically Ill Cancer Patients. Medicine (Baltimore). 2004 Nov;83(6):360–70.

27. Torpy JM, Campbell AD, Glass RM. Mechanical Ventilation. JAMA. 2010 Mar 3;303(9):902.
28. Young D, Harrison DA, Cuthbertson BH, Rowan K, TracMan Collaborators for the. Effect of Early vs Late Tracheostomy Placement on Survival in Patients Receiving Mechanical Ventilation: The TracMan Randomized Trial. JAMA. 2013 May 22;309(20):2121–9.
29. Roca O, Riera J, Torres F, Masclans JR. High-Flow Oxygen Therapy in Acute Respiratory Failure. Respir Care. 2010 Apr 1;55(4):408–13.
30. Wong AKI, Kamaleswaran R, Tabaie A, Reyna MA, Josef C, Robichaux C, et al. Prediction of Acute Respiratory Failure Requiring Advanced Respiratory Support in Advance of Interventions and Treatment: A Multivariable Prediction Model From Electronic Medical Record Data. Crit Care Explor. 2021 May 12;3(5):e0402.
31. Chen T, Guestrin C. XGBoost: A Scalable Tree Boosting System. Proc 22nd ACM SIGKDD Int Conf Knowl Discov Data Min. 2016 Aug 13;785–94.
32. Rusdah DA, Murfi H. XGBoost in handling missing values for life insurance risk prediction. SN Appl Sci. 2020 Jul 6;2(8):1336.
33. Ling CX, Huang J, Zhang H. AUC: A Better Measure than Accuracy in Comparing Learning Algorithms. In: Xiang Y, Chaib-draa B, editors. Advances in Artificial Intelligence. Berlin, Heidelberg: Springer; 2003. p. 329–41.
34. Rosset S. Model selection via the AUC. In: Proceedings of the twenty-first international conference on Machine learning [Internet]. New York, NY, USA: Association for Computing Machinery; 2004 [cited 2022 Mar 29]. p. 89. (ICML '04). Available from: https://doi.org/10.1145/1015330.1015400
35. Huang J, Ling CX. Using AUC and accuracy in evaluating learning algorithms. IEEE Trans Knowl Data Eng. 2005 Mar;17(3):299–310.
36. Moor M, Rieck B, Horn M, Jutzeler CR, Borgwardt K. Early Prediction of Sepsis in the ICU Using Machine Learning: A Systematic Review. Front Med. 2021;8:348.
37. Reyna MA, Josef CS, Jeter R, Shashikumar SP, Westover MB, Nemati S, et al. Early Prediction of Sepsis From Clinical Data: The PhysioNet/Computing in Cardiology Challenge 2019. Crit Care Med. 2020 Feb;48(2):210–7.
38. Subbe CP, Kruger M, Rutherford P, Gemmel L. Validation of a modified Early Warning Score in medical admissions. QJM Mon J Assoc Physicians. 2001 Oct;94(10):521–6.
39. Schermer T, Leenders J, in 't Veen H, van den Bosch W, Wissink A, Smeele I, et al. Pulse oximetry in family practice: indications and clinical observations in patients with COPD. Fam Pract. 2009 Dec 1;26(6):524–31.
40. Kane B, Decalmer S, O'Driscoll BR. Emergency oxygen therapy: from guideline to implementation. Breathe. 2013 Jun 1;9(4):246–53.
41. O'Driscoll BR, Howard LS, Davison AG. BTS guideline for emergency oxygen use in adult patients. Thorax. 2008 Oct 1;63(Suppl 6):vi1–68.
42. Burdick H, Lam C, Mataraso S, Siefkas A, Braden G, Dellinger RP, et al. Prediction of respiratory decompensation in Covid-19 patients using machine learning: The READY trial. Comput Biol Med. 2020 Sep;124:103949.
43. Bolourani S, Brenner M, Wang P, McGinn T, Hirsch JS, Barnaby D, et al. A Machine Learning Prediction Model of Respiratory Failure Within 48 Hours of Patient Admission for COVID-19: Model Development and Validation. J Med Internet Res [Internet]. 2021 Feb [cited 2022 Mar 14];23(2). Available from: https://www.ncbi.nlm.nih.gov/labs/pmc/articles/PMC7879728/
44. Lundberg SM, Lee SI. A Unified Approach to Interpreting Model Predictions.:10.
45. Goligher EC, Brochard LJ, Reid WD, Fan E, Saarela O, Slutsky AS, et al. Diaphragmatic myotrauma: a mediator of prolonged ventilation and poor patient outcomes in acute respiratory failure. Lancet Respir Med. 2019 Jan 1;7(1):90–8.
46. Beitler JR, Malhotra A, Thompson BT. Ventilator-Induced Lung Injury. Clin Chest Med. 2016 Dec;37(4):633–46.
47. Brochard L, Slutsky A, Pesenti A. Mechanical Ventilation to Minimize Progression of Lung Injury in Acute Respiratory Failure. Am J Respir Crit Care Med. 2017 Feb 15;195(4):438–42.

48. Wilson JG, Matthay MA. Mechanical Ventilation in Acute Hypoxemic Respiratory Failure: A Review of New Strategies for the Practicing Hospitalist. J Hosp Med Off Publ Soc Hosp Med. 2014 Jul;9(7):469–75.
49. Kim J, Chae M, Chang HJ, Kim YA, Park E. Predicting Cardiac Arrest and Respiratory Failure Using Feasible Artificial Intelligence with Simple Trajectories of Patient Data. J Clin Med. 2019 Aug 29;8(9):1336.
50. Brown R, McKelvey MC, Ryan S, Creane S, Linden D, Kidney JC, et al. The Impact of Aging in Acute Respiratory Distress Syndrome: A Clinical and Mechanistic Overview. Front Med [Internet]. 2020 [cited 2022 Apr 29];7. Available from: https://www.frontiersin.org/article/10.3389/fmed.2020.589553
51. Gowda S, Desai PB, Kulkarni SS, Hull VV, Math AAK, Vernekar SN. Markers of renal function tests. North Am J Med Sci. 2010 Apr;2(4):170–3.
52. Darmon M, Legrand M, Terzi N. Understanding the kidney during acute respiratory failure. Intensive Care Med. 2017 Aug 1;43(8):1144–7.
53. Li X, Ma X. Acute respiratory failure in COVID-19: is it "typical" ARDS? Crit Care. 2020 May 6;24:198.
54. Ting C, Aspal M, Vaishampayan N, Huang SK, Riemondy KA, Wang F, et al. Fatal COVID-19 and Non–COVID-19 Acute Respiratory Distress Syndrome Is Associated with Incomplete Alveolar Type 1 Epithelial Cell Differentiation from the Transitional State without Fibrosis. Am J Pathol. 2022 Mar 1;192(3):454–67.
55. Aslan A, Aslan C, Zolbanin NM, Jafari R. Acute respiratory distress syndrome in COVID-19: possible mechanisms and therapeutic management. Pneumonia. 2021 Dec 6;13(1):14.

# Figures

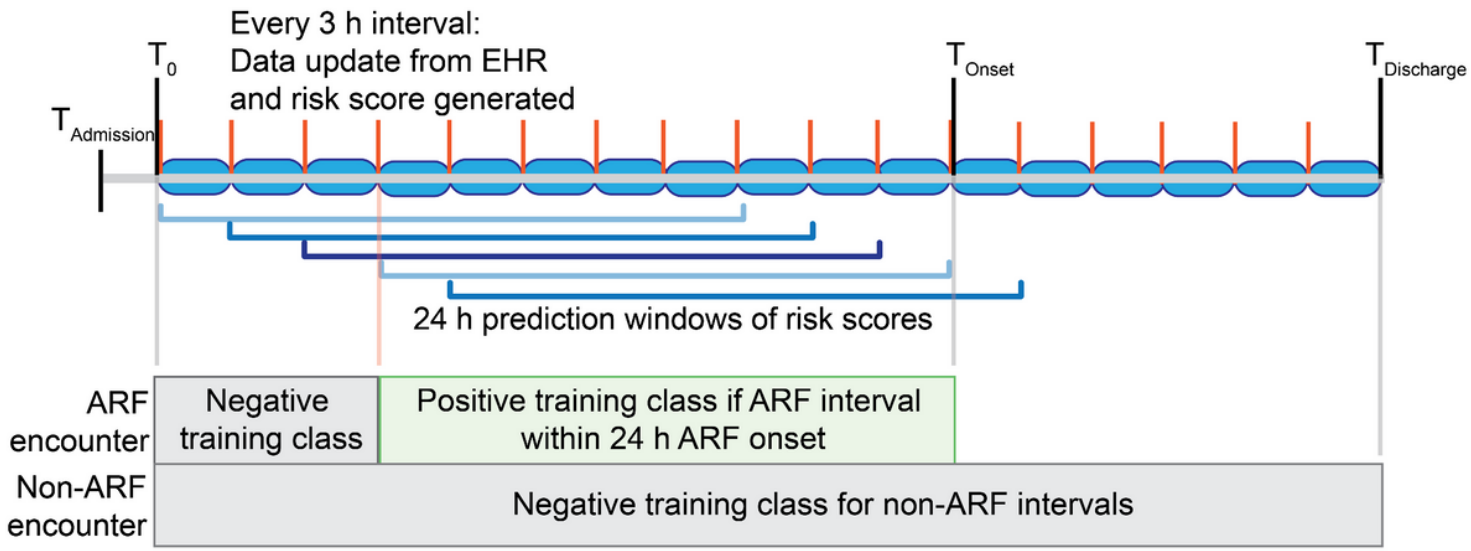


**Figure 1**

Training strategy for ARF and non-ARF encounters using 3-hour data intervals. $T_0$ denotes the time at least one measurement from each required feature is obtained. The orange mark indicates a 3-hour interval that culminates in a data update from EHR which generates a risk score. The risk score predicts the risk of ARF onset within 24 hours. The 3-hour data intervals for ARF encounters are classified as positive if ARF onset occurs within 24 hours, and a member of the negative class otherwise. Data intervals for non-ARF encounters are members of the negative training class. Acute respiratory failure (ARF), hour (h), time (T).

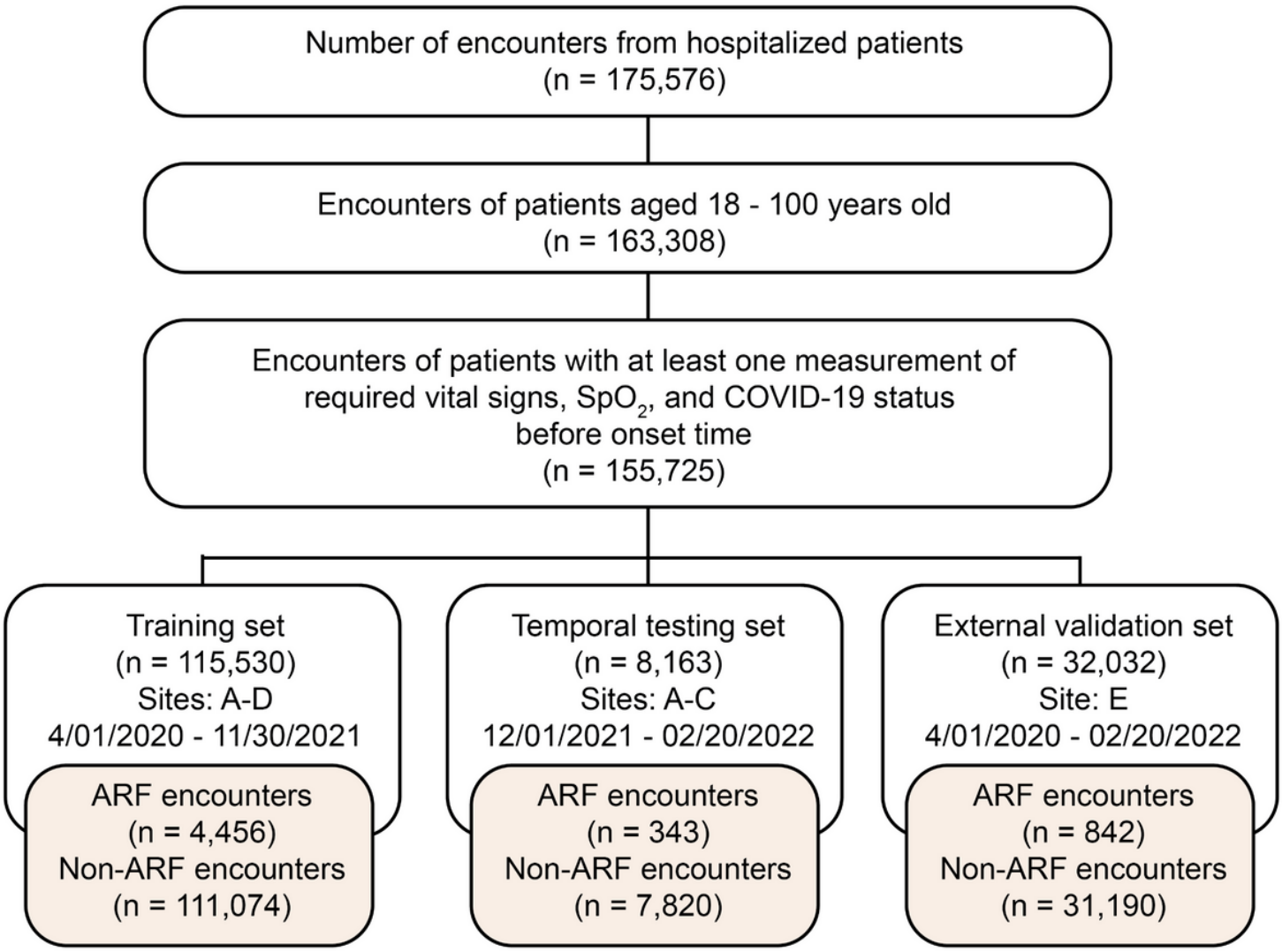


**Figure 2**

Flowchart for the inclusion criteria and number of encounters included in the analysis. Encounter data was obtained from the electronic health records of patients from five United States hospitals. Data from the training and temporal testing set did not overlap. ARF, acute respiratory failure; $SpO_2$, saturated oxygen level.

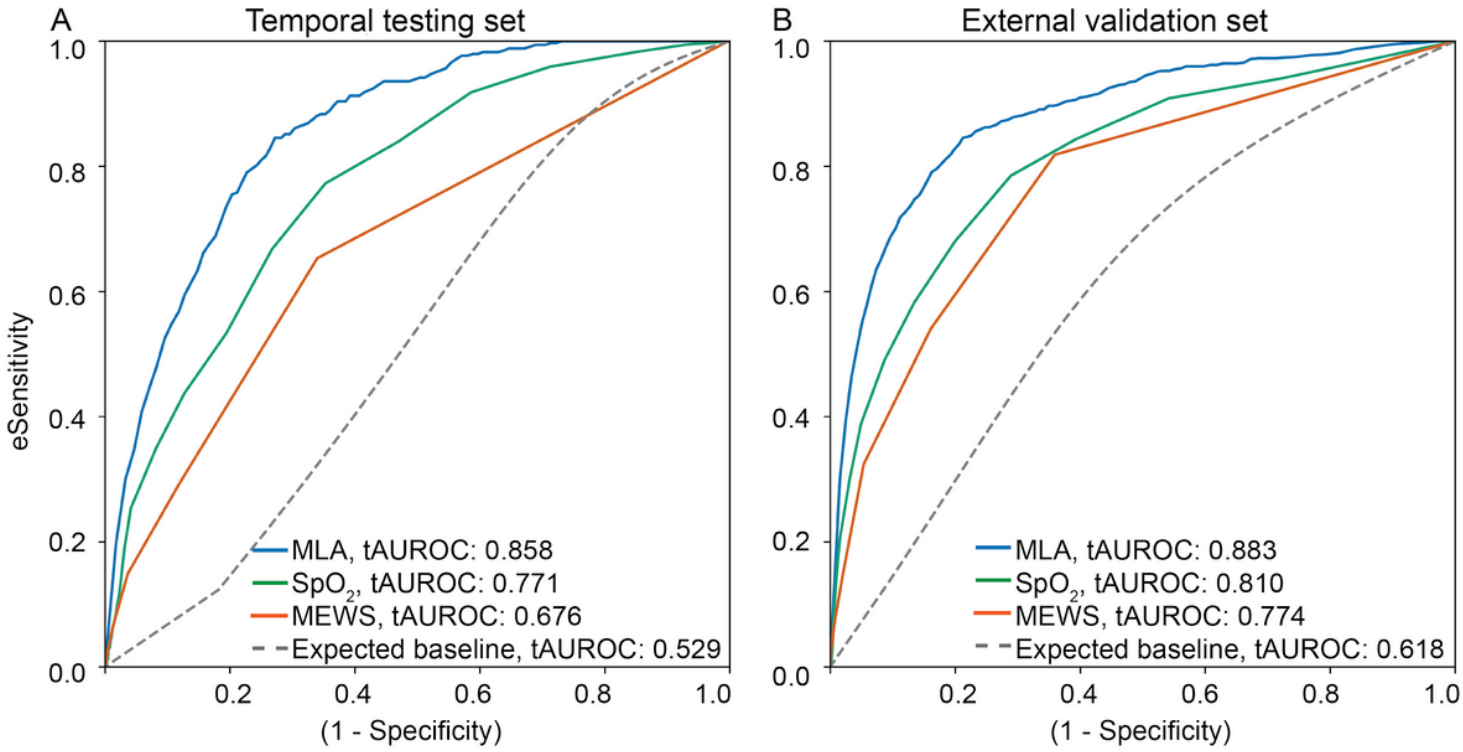


**Figure 3**

Time-sensitive receiver operating characteristic (tROC) curves for the evaluation of the early prediction of ARF machine learning algorithm (MLA) in comparison to oxygen saturation ($SpO_2$) and modified early warning score (MEWS) comparators on the temporal testing set (A) and external validation set (B). Due to the model producing random scores at each time point relative to the onset of ARF, the baseline random model does not have an area under the curve of 0.5 as for a typical ROC. Time-sensitive area under the receiver operating characteristic (tAUROC).

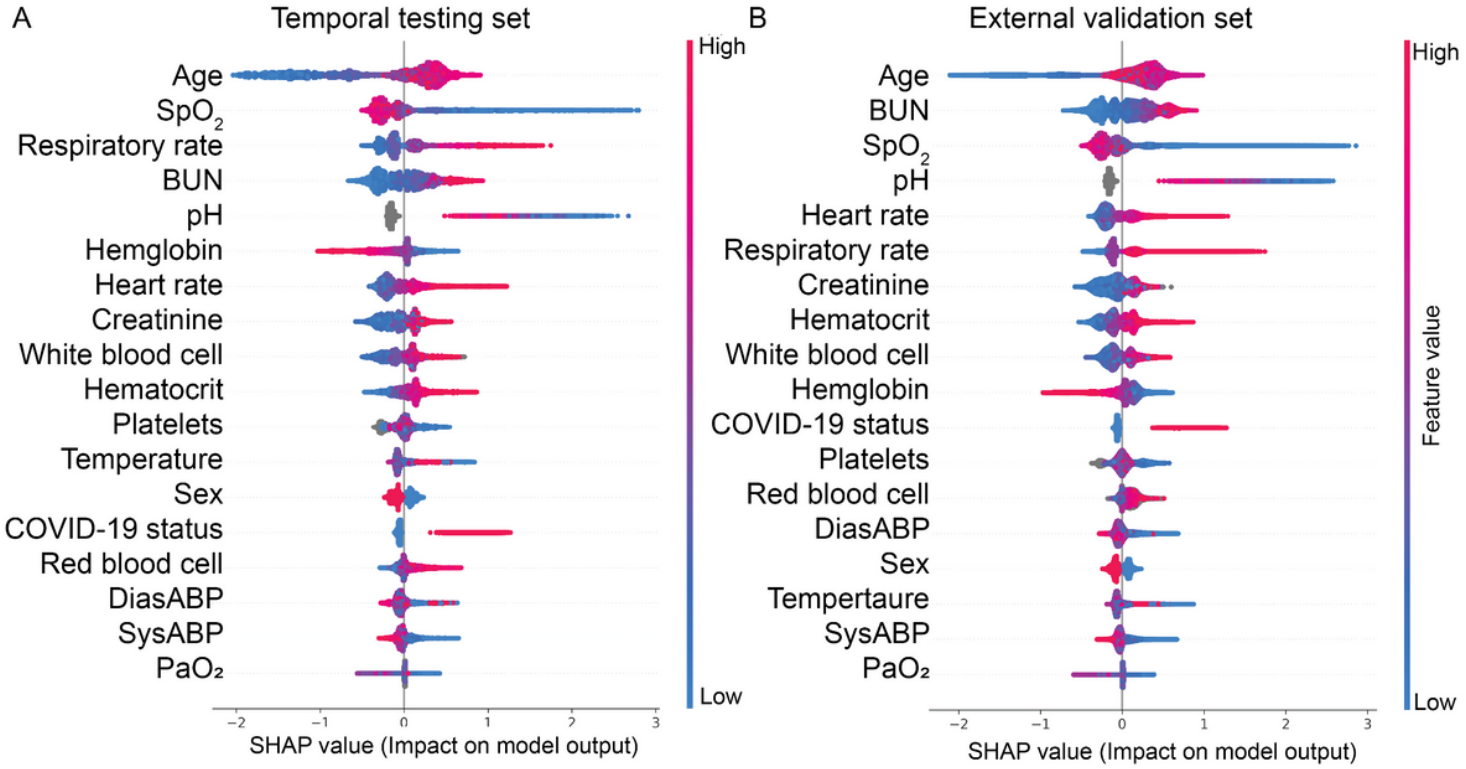


**Figure 4**

SHAP analysis of feature importance of the early ARF prediction machine learning algorithm on the temporal testing (A) and external validation (B) sets. The features are ranked from the top in order of highest to lowest importance. Red denotes a high feature value and blue indicates a low value. The impact of a feature on the model prediction is correlated by positive or negative SHAP values. Blood urea nitrogen (BUN); diastolic arterial blood pressure (DiasABP); oxygen saturation ($SpO_2$); partial pressure of oxygen ($PaO_2$); systolic arterial blood pressure (SysABP).

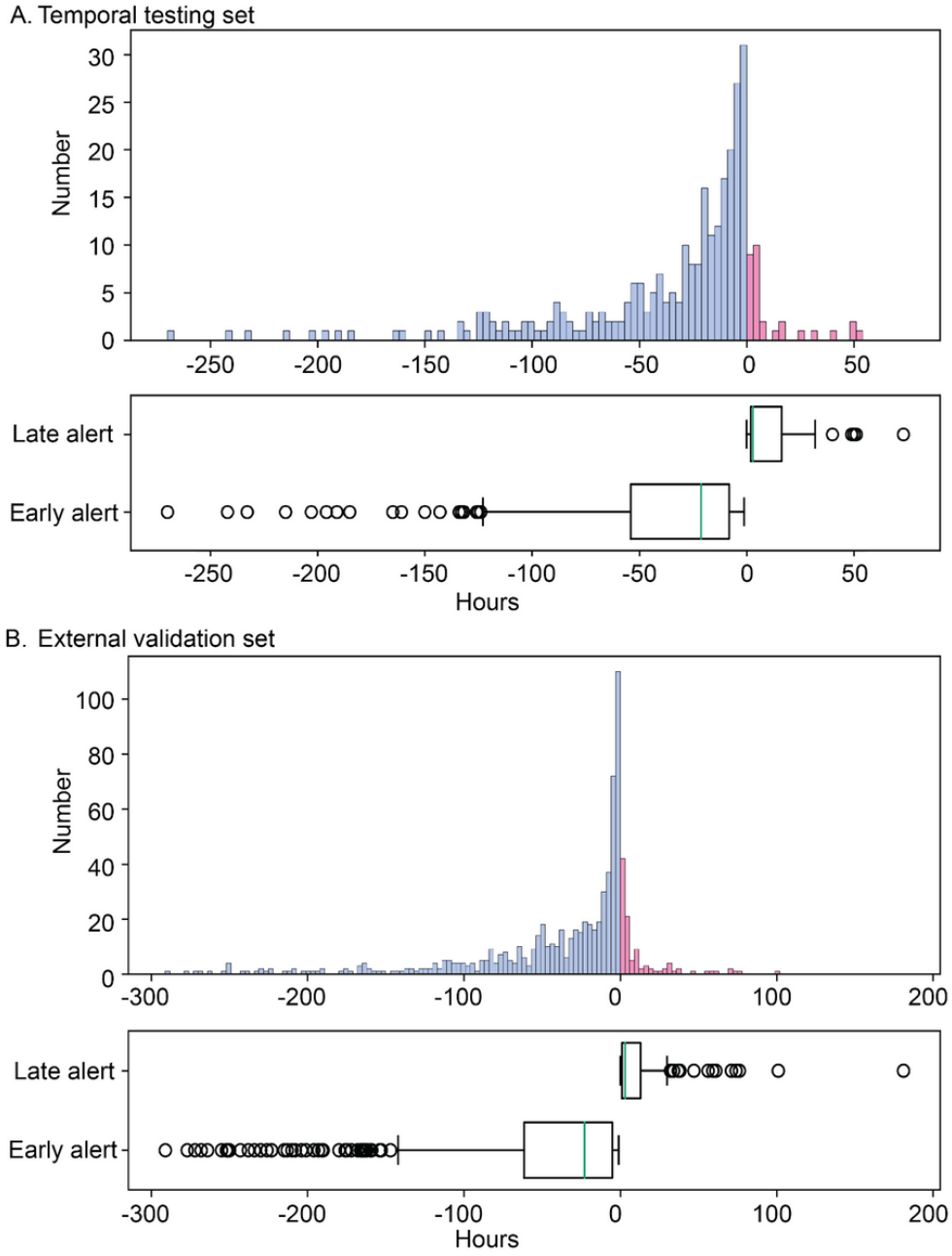


**Figure 5**

Alert timing histogram for acute respiratory failure (ARF) encounters that were alerted by the machine learning algorithm. The timing analysis on the temporal testing set (**A**) demonstrated 266 *early* alerts (left; blue bars) were generated prior to the onset of ARF, as well as 46 late alerts (right; red bars) that were produced after ARF onset. The median value (green line) for *early* alerts was 21.0 hours. On the external validation set (**B**), 630 *early* alerts (left; blue bars) were produced, as well as 105 alerts that were produced after ARF onset (right; red bars). The median value (green line) for *early* alerts was 23.0 hours. X-axis denotes the time in hours from the first alert produced in positive patients with ARF onset time = 0, and the y-axis is the number of alerts.

## Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryMaterials.docx