

UC Davis

UC Davis Previously Published Works

Title

Use of Machine Learning to Screen for Acute Respiratory Distress Syndrome Using Raw Ventilator Waveform Data

Permalink

<https://escholarship.org/uc/item/6x59518v>

Journal

Critical Care Explorations, 3(1)

ISSN

2639-8028

Authors

Rehm, Gregory B
Cortés-Puch, Irene
Kuhn, Brooks T
[et al.](#)

Publication Date

2021

DOI

10.1097/cce.0000000000000313

Copyright Information

This work is made available under the terms of a Creative Commons Attribution-NonCommercial-NoDerivatives License, available at <https://creativecommons.org/licenses/by-nc-nd/4.0/>

Peer reviewed

OPEN

Use of Machine Learning to Screen for Acute Respiratory Distress Syndrome Using Raw Ventilator Waveform Data

OBJECTIVES: To develop and characterize a machine learning algorithm to discriminate acute respiratory distress syndrome from other causes of respiratory failure using only ventilator waveform data.

DESIGN: Retrospective, observational cohort study.

SETTING: Academic medical center ICU.

PATIENTS: Adults admitted to the ICU requiring invasive mechanical ventilation, including 50 patients with acute respiratory distress syndrome and 50 patients with primary indications for mechanical ventilation other than hypoxemic respiratory failure.

INTERVENTIONS: None.

MEASUREMENTS AND MAIN RESULTS: Pressure and flow time series data from mechanical ventilation during the first 24-hours after meeting acute respiratory distress syndrome criteria (or first 24-hr of mechanical ventilation for non-acute respiratory distress syndrome patients) were processed to extract nine physiologic features. A random forest machine learning algorithm was trained to discriminate between the patients with and without acute respiratory distress syndrome. Model performance was assessed using the area under the receiver operating characteristic curve, sensitivity, specificity, positive predictive value, and negative predictive value. Analyses examined performance when the model was trained using data from the first 24 hours and tested using withheld data from either the first 24 hours (24/24 model) or 6 hours (24/6 model). Area under the receiver operating characteristic curve, sensitivity, specificity, positive predictive value, and negative predictive value were 0.88, 0.90, 0.71, 0.77, and 0.90 (24/24); and 0.89, 0.90, 0.75, 0.83, and 0.83 (24/6).

CONCLUSIONS: Use of machine learning and physiologic information derived from raw ventilator waveform data may enable acute respiratory distress syndrome screening at early time points after intubation. This approach, combined with traditional diagnostic criteria, could improve timely acute respiratory distress syndrome recognition and enable automated clinical decision support, especially in settings with limited availability of conventional diagnostic tests and electronic health records.

KEY WORDS: acute respiratory distress syndrome; classification; critical care; mechanical ventilation; population surveillance; respiratory failure

Acute respiratory distress syndrome (ARDS) is a severe form of hypoxemic respiratory failure present in up to 10% of ICU admissions and 25% of patients receiving mechanical ventilation (MV) (1). Patients with ARDS experience substantial morbidity and mortality, prolonged MV,

Gregory B. Rehm, MS¹

Irene Cortés-Puch, MD²

Brooks T. Kuhn, MD, MAS²

Jimmy Nguyen, RRT³

Sarina A. Fazio, PhD, RN²

Michael A. Johnson, MD, PhD⁴

Nicholas R. Anderson, PhD⁵

Chen-Nee Chuah, PhD⁶

Jason Y. Adams, MD, MS²

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.0000000000000313

high hospital-associated costs, and long-term physical and psychologic dysfunctions (1, 2). Poor outcomes in ARDS are associated with delayed and missed diagnosis, and suboptimal use of evidence-based therapies even by subspecialty-trained clinicians (1–4). Diagnostic criteria require arterial blood gas (ABG) measurement of the ratio of P_{aO_2} to F_{iO_2} , referred to as “P/F,” and the presence of bilateral opacities on chest imaging, which may contribute to delayed and/or missed diagnosis given decreasing use of ABGs in critical care and poor interrater agreement in chest x-ray (CXR) interpretation (5–9). Even when the aforementioned clinical variables are available and ARDS criteria are met, clinicians identified only 65.3% of moderate and 78.5% of severe ARDS cases at any point during ICU admission, and only 34% of cases on the first day (1).

Challenges associated with early identification of ARDS have led to the development of automated ARDS screening systems. Early examples of these so-called ARDS “sniffer” systems used rule-based processing of keywords extracted from CXR reports and screening of ABG data for qualifying P/F values (10, 11), whereas more recent studies have used natural language processing and machine learning (ML) algorithms (12). Although these studies demonstrated the potential of automated systems to improve the accuracy and timeliness of ARDS diagnosis, each approach was dependent on the availability of laboratory and radiographic data, local radiologist practices, and information technology that may not be present in all healthcare settings. When the generalizability of ARDS sniffers was tested in new patient populations, algorithm specificity declined substantially (13, 14).

To address the limitations of data timeliness and availability, and the challenges of extracting information from imaging reports, we hypothesized that raw ventilator waveform data (VWD) could be used to screen newly intubated patients early in the course of moderate-severe ARDS. VWD is particularly appealing for syndrome surveillance as the data contain quantitative physiologic information and are continuously available from the start of MV. Previous studies using both rule-based and ML-based methods have shown that VWD can be used to automate the assessment of patient-ventilator asynchrony and exposure to excessive tidal volumes (TVs), and that VWD may be useful to monitor progression of ARDS (15–23). However,

VWD has not yet been studied in the context of ARDS screening. We thus specifically hypothesized that an ML model, using only physiologic information extracted from raw VWD, would be able to discriminate between patients with and without ARDS at early time points in MV without the need for CXR, ABG, or other electronic health record (EHR)-derived data.

MATERIALS AND METHODS

Cohort Selection

All patient data were obtained as part of a prospective, Institutional Review Board approved study collecting raw VWD from mechanically ventilated adults admitted to the Medical ICU at the UC Davis Medical Center. Three clinicians (I.C.P., B.T.K., and J.Y.A.) performed retrospective chart review to identify the cause of respiratory failure in subjects from the VWD study cohort enrolled between 2015 and 2019. Subjects were split into two cohorts: 1) patients with confirmed moderate or severe ARDS diagnosed using Berlin consensus criteria within 7 days of intubation (5) and 2) patients with no suspicion of ARDS during their MV course, to avoid phenotypic ambiguity. Patients with chronic obstructive pulmonary disease and/or asthma were excluded from the ARDS patient cohort to minimize the risk of misclassifying ARDS as a result of concurrent non-ARDS acute or nonacute chronic lung disease-associated hypoxemia. Causes of ARDS and indications for MV in the non-ARDS cohort are shown in **Table 1**, and additional clinical information such as primary ventilator mode, depth of sedation, use of neuromuscular blockade, and rates of two common asynchronies are shown in **Supplemental Digital Content Table 2** (<http://links.lww.com/CCX/A480>) and **Supplemental Digital Content Figure 2** (<http://links.lww.com/CCX/A479>). Both cohorts required at least 1 hour of VWD collected in the first 24 hours after ARDS criteria were first met or after the start of MV (**Supplemental Digital Content Table 1**, <http://links.lww.com/CCX/A480>, and **Supplemental Digital Content Fig. 1**, <http://links.lww.com/CCX/A479>). All cases meeting study inclusion criteria were reviewed by two clinicians and only cases that were unambiguously considered to have ARDS or to not have ARDS were included in the study cohort. No sample size calculation was conducted for this study; however, sample size was guided by the range of cohort sizes in previous studies of VWD analysis (15–19) and

TABLE 1.
Clinical Characteristics of Study Subjects

| | ARDS (n = 50) | Non-ARDS (n = 50) |
|--|----------------------|--------------------------|
| Age (median [IQR]) | 57 (38–65) | 58 (49–67) |
| Female (n [%]) | 13 (26) | 23 (46) |
| Body mass index (median [IQR]) | 26.4 (22.3–33.8) | 25.9 (22.1–28.7) |
| Obstructive lung disease (n [%]) | | |
| COPD | 0 (0) | 12 (24) |
| Asthma | 0 (0) | 5 (10) |
| Reason for ICU admission (n [%]) | | |
| Acute hypoxemic respiratory failure | 24 (48) | – |
| COPD/asthma exacerbation | – | 17 (34) |
| Sepsis | 11 (22) | – |
| Metabolic encephalopathy/drug overdose | 2 (4) | 15 (30) |
| Airway edema/anaphylaxis | – | 5 (10) |
| Stroke | – | 4 (8) |
| Cardiac arrest | 9 (18) | 3 (6) |
| Heart failure | – | 2 (4) |
| Upper gastrointestinal bleeding | – | 2 (4) |
| Trauma/surgery | 3 (6) | 2 (4) |
| Pancreatitis | 1 (2) | – |
| Sequential Organ Failure Assessment score (median [IQR]) | 13 (10–16) | 7.5 (5–10) |
| Days from intubation to Berlin criteria (median [IQR]) | 0.1 (0.0–0.2) | – |
| Median Pao ₂ /Fio ₂ first 24 hr (median [IQR]) | 176 (134–210) | 318 (267–423) |
| Worst Pao ₂ /Fio ₂ 24 hr (median [IQR]) | 108 (66–137) | 278 (147–385) |
| ARDS insult type (n [%]) | | |
| Pneumonia | 18 (36) | – |
| Aspiration | 14 (28) | – |
| Nonpulmonary sepsis | 10 (20) | – |
| Trauma | 2 (4) | – |
| Diffuse alveolar hemorrhage | 2 (4) | – |
| Pancreatitis | 1 (2) | – |
| Other | 3 (6) | – |
| Hospital length of stay (median [IQR]) | 13.3 (6.6–25.4) | 7.0 (4.2–13.4) |
| Hospital mortality (n [%]) | 24 (48) | 10 (20) |
| Ventilator-free days in 28 d (median [IQR]) | 6.6 (0–23.0) | 25.3 (10.6–26.9) |

ARDS = acute respiratory distress syndrome, COPD = chronic obstructive pulmonary disease, IQR = interquartile range.

to achieve a balanced dataset for ML model development as standard ML algorithms are biased toward the majority class, resulting in a higher misclassification rate for the minority class (24).

VWD Acquisition and Featurization

We used our ventMAP software platform (18) to extract nine physiologic features from raw VWD, representing pressure and flow, sampled at 50 Hz, obtained from Puritan-Bennett model 840 ventilators (25). Features were extracted and aimed to capture relevant respiratory pathophysiology, while avoiding features that might strongly correlate with ARDS management such as TV or positive end-expiratory pressure (**Supplemental Digital Content Table 3**, <http://links.lww.com/CCX/A480>). Observations for each subject were derived by taking the median value of each feature across windows of 100 consecutive VWD breaths (approximately 5 min), with the 100-breath window size based on empirical sensitivity analysis. We processed all available VWD in the 24 hours after Berlin criteria were first met and 24 hours after the start of MV for patients with and without ARDS, respectively (**Fig. 1A**). However, not all patients had 24 hours of data based on variable start times of data acquisition. Each observation feature vector was tagged with a subject identifier and clinical class label of ARDS versus non-ARDS. Observations were excluded if a feature in the observation window met any of following: 1) not a number or an infinite value, 2) more than 50% of expected breaths in a window were missing, or 3) window start time was prior to charted MV start and end times in the EHR.

Machine Learning Model Development

We evaluated seven algorithms using the Python scikit-learn software library (**Supplemental Digital Content Table 4**, <http://links.lww.com/CCX/A480>) (26). Despite comparable performance across algorithms, we chose the random forest (RF) algorithm for further model development and testing based on its resistance to overfitting and tolerance to outliers (27). Given our small sample size, we evaluated model performance using k -fold cross validation ($k = 5$), a 70/30 holdout split, and bootstrapping. For the five-fold cross validation, 80 subjects were used for training in each of the five k -folds, and 20 were used for validation, with no overlapping data between the training

and validation sets in each fold (**Supplemental Digital Content Fig. 3**, <http://links.lww.com/CCX/A479>). For the 70/30 holdout split, we randomly selected 70 subjects for model training, and withheld 30 for final model validation. For bootstrapping, 100 bootstrapping runs were performed. In each run, 80 patients were randomly selected with replacement for training and the remaining patients were used for validation, with performance averaged over 100 bootstraps. We performed feature selection for each model using the chi-square and Gini selection methods. Model hyperparameters were selected using a Python grid search (**Supplemental Digital Content Table 5**, <http://links.lww.com/CCX/A480>) (26). Because feature importance was comparable across both methods, we used sequential feature selection to maximize the area under the receiver operating curve (AUC) with chi-square for all models (**Supplemental Digital Content Tables 6–8**, <http://links.lww.com/CCX/A480> and **Supplemental Digital Content Fig. 4**, <http://links.lww.com/CCX/A479>).

ARDS-screening ML models were developed using a two-step process. First, we trained the model to classify all individual 100-breath windows from the training set as either ARDS or non-ARDS (**Fig. 1B**). We then determined patient-level model performance by attributing all breath window predictions from the validation sets to each subject and assigning the subject class as ARDS or non-ARDS using a specific threshold for the percentage of individual windows classified as ARDS in any given time bin (**Fig. 1C**). We examined ML model performance to screen for ARDS using either 24 or 6 hours of VWD. We trained the 24/24 model using the first 24 hours of available VWD in the training set and then validated using the first 24 hours of available VWD from the validation set (24/24 model; $n = 100$). Our second model, the 24/6 model, was trained using the first 24 hours of available data but was validated using data available in the first 6 hours (24/6 model; $n = 70$). In both the 24/24 and 24/6 models, all available VWD were used within the specified time frames after Berlin criteria were first met or after the start of MV for ARDS and non-ARDS subjects, respectively.

Model performance was assessed using AUC, sensitivity, specificity, positive predictive value, and negative predictive value. Performance was compared using a simple majority voting threshold (e.g., more than

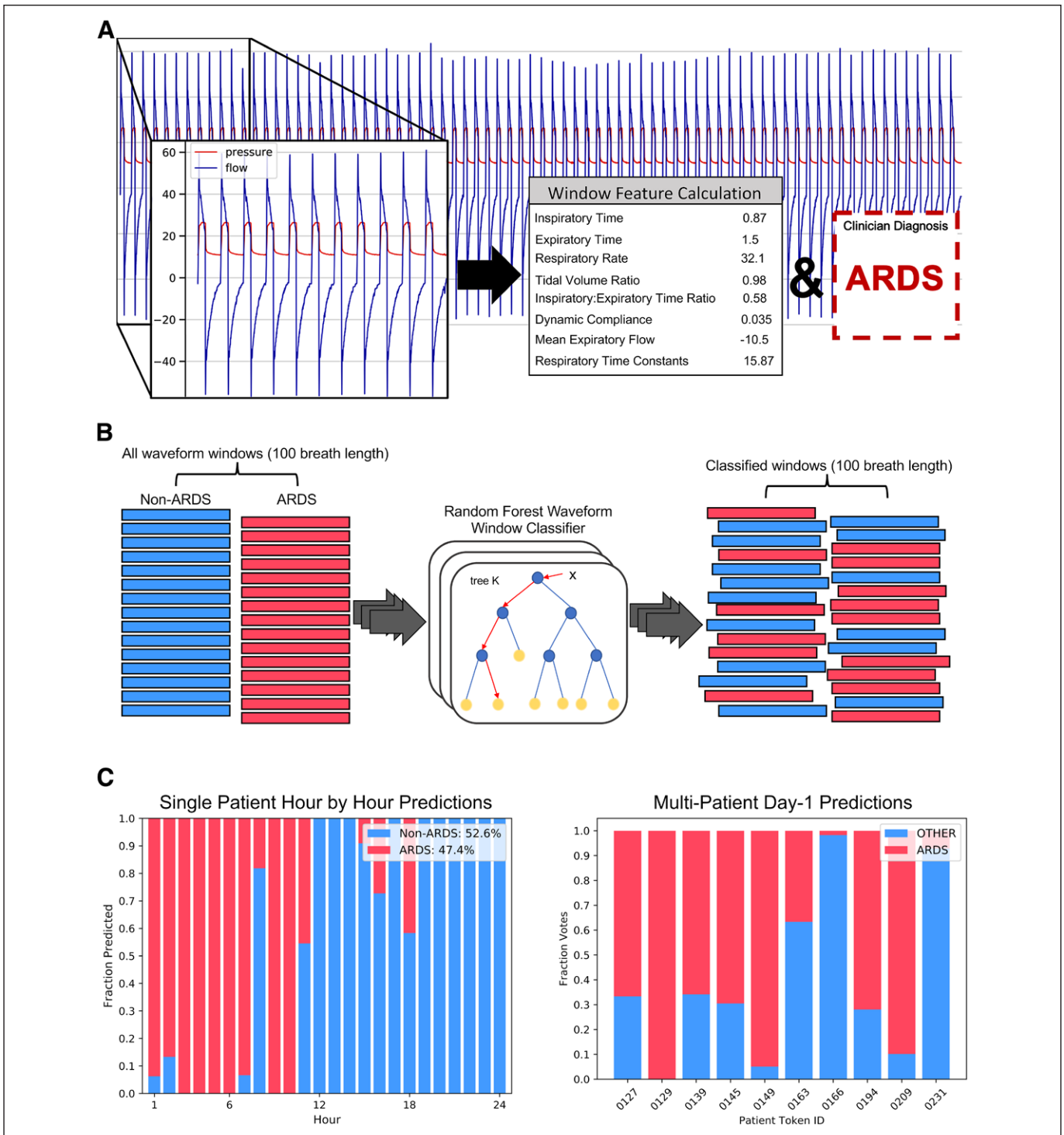


Figure 1. Visual overview of data processing and classifier model development. **A**, Ventilator waveform data from each subject were divided into consecutive 100-breath observation windows. Physiologic features were calculated for each breath in a window, and median values were used to represent the entire window. Each window was labeled as acute respiratory distress syndrome (ARDS) or non-ARDS and tagged with a subject identifier. **B**, Feature vectors for each labeled window were fed to a supervised machine learning algorithm for training and evaluation. **C**, Classified windows were aggregated at the patient level to allow threshold-based, patient-level predictions to be made based on the percentage of ARDS and non-ARDS windows within any given time period (e.g., 24 hr).

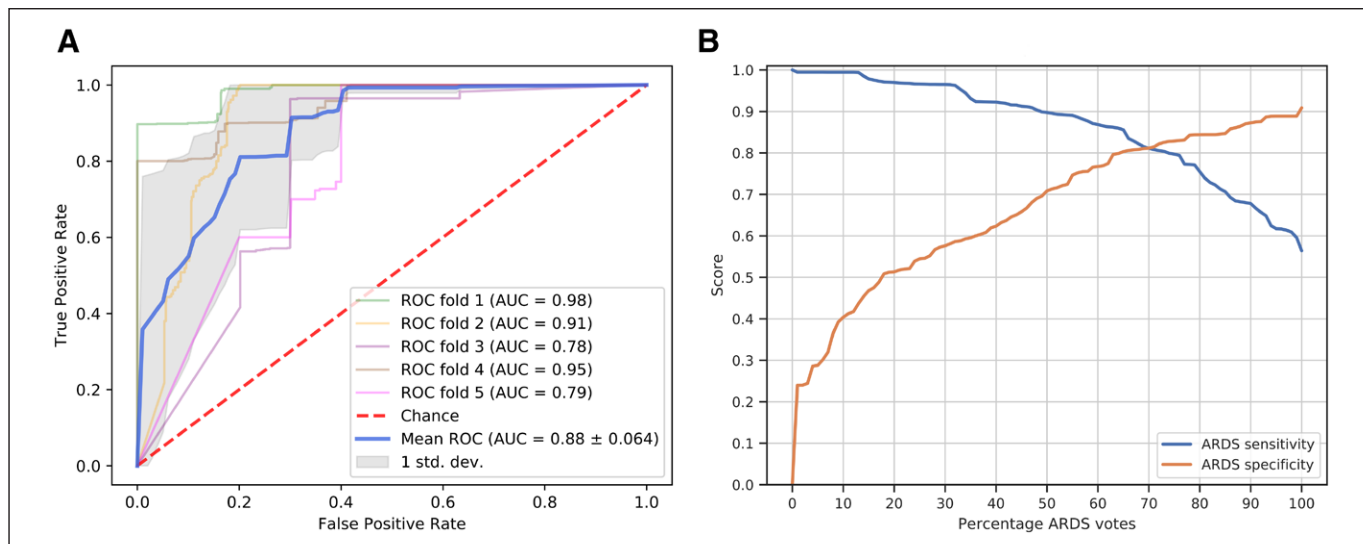


Figure 2. Performance characteristics of the train 24-/test 24-hr model (24/24). **A**, Receiver operating characteristic (ROC) curves for individual k -folds in the 24/24 five-fold cross validation model. Mean area under the ROC (area under the receiver operating characteristic curve [AUC]) across all k -folds is shown in *blue* (95% CI displayed in figure legend). **B**, Sensitivity and specificity of acute respiratory distress syndrome (ARDS) detection change as the voting threshold required to classify ARDS in the first 24 hr increases.

50% of 100-breath windows were classified as ARDS in a given time period) and across a range of voting threshold deciles between 0% and 100%.

RESULTS

A total of 100 adult mechanically ventilated patients were included in the study, including 50 with ARDS and 50 without evidence of ARDS during the course of MV. Table 1 provides demographic, clinical, and physiologic characteristics of subjects. We analyzed a median of 21.2 hours of VWD per subject from ARDS patients and 13.3 hours of VWD from non-ARDS patients, representing 19,777 100-breath window observations. The dataset contained a total of 2,020,556 breaths, with 1,331,285 breaths from patients with ARDS and 689,271 from patients without ARDS.

Performance of our primary ML model discriminating between ARDS and non-ARDS cases using the first 24 hours of VWD (24/24 model) is shown in **Figure 2A** and **Table 2**. For our main analyses, we used a simple majority voting scheme to determine patient-level predictions. Thus, if 51% or more observations from a patient were classified as ARDS, the patient was classified as ARDS. Using this voting threshold, the 24/24 VWD model was able to discriminate between the ARDS and non-ARDS subjects with a mean AUC across all five k -folds of 0.88 (95% CI, 0.816–0.944). Discriminative performance

was similar in the 70/30 holdout and bootstrapping experiments (**Supplemental Digital Content Tables 9 and 10**, <http://links.lww.com/CCX/A480>). **Figure 2B** and **Table 3** show how model sensitivity and specificity varied by changing the threshold used to classify ARDS across the range of prediction thresholds from 0% to 100% prediction votes, and at specific threshold deciles from 10% to 100%, respectively.

Our second ML model explored the ability of an ML algorithm trained on the first 24 hours of VWD to differentiate between ARDS and non-ARDS in a validation set using only the first 6 hours of VWD after meeting Berlin criteria or starting MV for the ARDS and the non-ARDS cohorts, respectively (24/6 model). Discriminative performance in this 24/6 model was comparable with the 24/24 model with AUCs of 0.89 (95% CI, 0.817–0.963) and 0.88 (95% CI, 0.816–0.944), respectively, using five-fold cross validation.

DISCUSSION

We developed an automated ARDS screening algorithm that can detect potential cases of moderate-severe ARDS early in the course of MV without need for CXR, ABG, or other EHR-derived data. Using ML techniques and physiologic features derived from raw VWD, our model demonstrated robust discriminative performance for detecting ARDS in the first 24 hours after meeting Berlin criteria that were reproducible

TABLE 2.**Model Performance Statistics for Both Train 24-/Test 24-hr (24/24) and Train 24-/Test 6-hr (24/6) Models**

| Model | Train/Test Split (<i>n</i>) | <i>k</i> -Fold Number | Sensitivity | Specificity | Positive Predictive Value | Negative Predictive Value | Area Under the Curve |
|----------------------|-------------------------------|------------------------------|--------------|--------------|---------------------------|---------------------------|----------------------|
| Train 24/ Test 24 | 80/20 | 1 | 1.0 | 0.73 | 0.79 | 1.0 | 0.98 |
| – | – | 2 | 1.0 | 0.79 | 0.83 | 1.0 | 0.92 |
| – | – | 3 | 0.70 | 0.70 | 0.69 | 0.70 | 0.78 |
| – | – | 4 | 0.80 | 0.91 | 0.90 | 0.82 | 0.94 |
| – | – | 5 | 1.0 | 0.44 | 0.64 | 1.0 | 0.79 |
| – | Not applicable | Mean of five <i>k</i> -folds | 0.90 ± 0.059 | 0.71 ± 0.089 | 0.77 ± 0.082 | 0.90 ± 0.059 | 0.88 ± 0.064 |
| Train 24/ Test 6 | 80/14 | Mean of five <i>k</i> -folds | 0.90 ± 0.07 | 0.75 ± 0.101 | 0.83 ± 0.088 | 0.83 ± 0.088 | 0.89 ± 0.073 |

Mean (with 95% CIs) performance across all five *k*-folds is shown for both models, and results of individual *k*-folds are displayed for the 24/24 model to illustrate the spectrum of performance variability. Note that only 70 subjects had ventilator waveform data available in the first 6 hr, resulting in a smaller sample size for the test cohort in the 24/6 model (see Supplemental Digital Content Table 8, <http://links.lww.com/CCX/A480>, for individual *k*-fold results of the 24/6 model).

TABLE 3.**Performance Characteristics of the Train 24/Test 24-hr (24/24) Model for Detection of Acute Respiratory Distress Syndrome Across Deciles of Voting Thresholds, Illustrating the Tunable Nature of Our Two-Step Acute Respiratory Distress Syndrome Classification Methodology**

| % Acute Respiratory Distress Syndrome Votes in First 24 hr | Sensitivity | Specificity | Positive Predictive Value | Negative Predictive Value |
|--|--------------|--------------|---------------------------|---------------------------|
| 10 | 0.99 ± 0.02 | 0.4 ± 0.096 | 0.63 ± 0.095 | 0.99 ± 0.02 |
| 20 | 0.97 ± 0.033 | 0.51 ± 0.098 | 0.68 ± 0.091 | 0.96 ± 0.038 |
| 30 | 0.96 ± 0.038 | 0.58 ± 0.097 | 0.71 ± 0.089 | 0.96 ± 0.038 |
| 40 | 0.92 ± 0.053 | 0.63 ± 0.095 | 0.74 ± 0.086 | 0.92 ± 0.053 |
| 50 | 0.9 ± 0.059 | 0.71 ± 0.089 | 0.77 ± 0.082 | 0.91 ± 0.056 |
| 60 | 0.87 ± 0.066 | 0.77 ± 0.082 | 0.81 ± 0.077 | 0.87 ± 0.066 |
| 70 | 0.81 ± 0.077 | 0.81 ± 0.077 | 0.83 ± 0.074 | 0.83 ± 0.074 |
| 80 | 0.75 ± 0.085 | 0.85 ± 0.07 | 0.85 ± 0.07 | 0.79 ± 0.08 |
| 90 | 0.68 ± 0.091 | 0.87 ± 0.066 | 0.86 ± 0.068 | 0.74 ± 0.086 |
| 100 | 0.57 ± 0.097 | 0.91 ± 0.056 | 0.86 ± 0.068 | 0.69 ± 0.091 |

across a variety of experimental conditions. We further showed that our ARDS detection model could identify potential ARDS cases as early as 6 hours after Berlin criteria were first documented and that our model architecture enabled adjustment of model performance according to desired levels of sensitivity and specificity.

Despite intensive research into the etiology, diagnosis, and treatment of ARDS, multiple studies have shown that bedside providers continue to underrecognize the syndrome. In the largest multinational prospective cohort study to date, recognition of ARDS occurred in only 34% of patients on the first day when Berlin diagnostic criteria were present, and ever in only 60% of patients. Even when providers were prompted with the question “Did the patient have ARDS at any stage of their ICU stay?,” 34.7% of patients with moderate and 21.5% with severe ARDS were never recognized at any time while in intensive care (1). Reasons for delayed or failed diagnosis remain incompletely understood, but underrecognition is not likely the result of subtle clinical findings, since 88% of patients already met Berlin criteria on day 1 of hypoxemic respiratory failure in LUNG SAFE and in 76% of patients at the time of intubation in the LOTUS-FRUIT study (1, 28). Underdiagnosis has also been associated with suboptimal care delivery. In this regard, studies have demonstrated repeatedly that clinicians, operating unassisted by decision support, consistently fail to apply evidence-based therapies (1, 4, 28–30), whereas at least one study has shown that clinical decision support driven by automated ARDS screening can decrease the delivery of potentially injurious MV (31). Collectively, these studies demonstrate a clear need for improved ARDS screening strategies.

Our results expand on previous research of automated ARDS screening “sniffer” systems. The original ARDS systems, developed in parallel at two U.S. institutions, used rule-based algorithms combining keyword searching of CXR reports and processing of ABG data to screen for ARDS and alert clinicians in near real-time (10, 11). These initial studies reported excellent diagnostic performance; however, specificity decreased substantially when they were externally validated at a different institution (13), underscoring the challenges of generalizing algorithms that depend on local practice and documentation patterns. Since the initial ARDS sniffers were developed, at least six additional ARDS detection tools have been described and

validated in single institutions, all using EHR-based data and/or imaging (12). Most of these second-generation sniffers have used ML approaches in an attempt to address the challenges of rule-based algorithms. Four have used ML techniques based on natural language processing and text mining of CXR reports (14, 32, 33) or image processing and feature extraction from CXR images (34). Two studies did not incorporate radiographic data and were based only on clinical data extracted from the medical and surgical history, charted vital signs, laboratory results, ventilatory settings, and medication use (35, 36). Most of these recent ARDS sniffer tools reported moderate to excellent diagnostic performance locally; however, none have been validated externally. Potential barriers to widespread usability of existing ARDS detection systems include dependence on local practice patterns of EHR adoption, documentation and ordering, and the requirement that clinicians document accurately and order tests in a timely manner.

To address these limitations, our methods differed from previous research in several notable ways. Our use of ML with VWD-derived features may overcome some limitations of previous feature extraction methods by capturing the physiologic signatures present in waveforms instead of relying on EHR or imaging data alone. Because VWD are generated from the start of MV, our methods enable continuous patient monitoring and may allow for more timely identification of potential ARDS cases, independent of ordering or documentation, which was suggested by our finding that ARDS could be detected as early as 6 hours after Berlin criteria were first met. As access to physiologic waveform data becomes more common, our exclusive use of VWD may also extend automated ARDS screening to resource-constrained care environments such as community and rural hospitals lacking well-developed EHRs, and in developing nations, battlefields and disaster relief zones where tests required to fulfill Berlin criteria may be in short supply or unavailable (37). Use of ventilator waveform analysis may thus improve both the timeliness and the ability to apply automated ARDS screening in diverse settings, which are particularly important since delayed or missed diagnosis is thought to be a major contributor to suboptimal implementation of evidence-based therapies for ARDS (1, 3, 4, 30).

In addition to extending previous research into developing automated ARDS screening systems, our work

further demonstrates the potential value of ML in the analysis of large volumes of untapped streaming physiologic waveform data generated from patient-monitoring devices in the ICU. The use of patient-derived physiologic data has gained increasing attention in recent years as the availability of both high-volume, high-sampling rate data types, and advanced computing power has become more commonplace. In this regard, automated processing of VWD has been demonstrated by multiple investigators in the study of patient-ventilator asynchrony (16–18) and several groups have shown the potential to computationally extract physiologic features from VWD including data derived from animal models of ARDS pertaining to airway resistance and respiratory system compliance (21–23). Sottile et al (19) and Rehm et al (20) have further investigated the ability to use ML to detect common types of patient-ventilator asynchrony without the need to explicitly code rule-based, expert systems, illustrating the ability of ML algorithms to learn relevant knowledge from the physiologic information embedded in raw VWD.

Our work also fits into a broader context of recent research using ML and sensor-derived physiologic data to develop so-called digital biomarkers to screen for and monitor diseases, improve disease phenotyping, and predict clinical trajectories (38). Recent studies in critical care have demonstrated the potential of digital biomarker signatures including the use of convolutional deep neural networks to process electrocardiogram waveforms to screen for hyperkalemia (39) and detect arrhythmias (40), and the use of continuous electroencephalography waveforms and deep learning to predict neurologic outcome after cardiac arrest (41). Within this framework, our results suggest the potential of ML and VWD to generate digital biomarker signatures of ARDS, either alone or in combination with conventional biomarkers (42), to aid clinicians in early detection, monitoring ARDS progression, and prognostication of patient outcomes.

Our study has a number of limitations that should be addressed in future studies. First, our study was limited to a single academic medical center, which could affect model generalizability despite our exclusive use of quantitative physiologic data (13). Second, our limited sample size may have resulted in model overfitting. We attempted to address this issue by using the RF algorithm, which may be inherently more resistant

to overfitting (27), and several different frameworks for model validation. Although most prior studies using VWD have been similarly limited in size (16–19), research on larger cohorts will be necessary to understand the full limitations of waveform-based ARDS screening. Similarly, our subject selection was intentionally biased to ensure phenotypic separation between ARDS and non-ARDS subjects to test the hypothesis that VWD and ML could be used to discriminate between clear phenotypes. Our cohort was comprised mostly of moderate-severe persistent ARDS that was present on intubation, and it is unclear how our model would perform in late-onset ARDS, mild ARDS patients, rapid resolvers (43), those with an uncertain diagnosis when Berlin criteria are first met, or in those with preexisting chronic lung disease (6, 36, 44). Third, we focused on clinician-driven feature extraction and one ML algorithm. It is possible that the use of other input features, including nonwaveform EHR-derived features or algorithms capable of end-to-end model development and automated featurization such as deep learning (45) may have improved performance. Fourth, although VWD are ubiquitous at the bedside, widespread access to these data for research purposes remains a challenge at present. Finally, studies aimed at developing ARDS classifiers, including ours, are limited by the inherent imprecision of the Berlin criteria (6, 46). Recognizing this fundamental limitation, our study focused on developing a tunable screening algorithm rather than one aimed at diagnosis. Development of ARDS classifiers that generalize well and are trusted by clinicians will require additional study with larger, more heterogeneous populations, may require improved methods of class assignment, such as advanced imaging and physiologic or digital biomarkers (29, 47, 48), and will ultimately require external validation followed by thoughtful integration into decision support workflows to realize intended patient benefits.

CONCLUSIONS

We report the performance of an automated, ML-based ARDS screening algorithm that can detect ARDS with strong discrimination performance within the first 24 hours after Berlin criteria are first met, without the need for CXR, ABG, or other EHR-derived data. Our focus on feature extraction exclusively from VWD suggests that this approach may enable ARDS screening very early after intubation and nearly continuously,

which may result in decreased time to recognition, improved generalization to other centers, and may enable screening in resource-constrained settings where ABG, radiographic testing, and critical care expertise may be unavailable or scarce. Although our results represent a first proof of concept that digital biomarkers derived from physiologic monitoring data can be used for ARDS detection, additional research is needed to determine how broadly such methods can be applied, how best to incorporate them into traditional Berlin criteria-based diagnostic work flows, and how they might compliment EHR and biochemical approaches to clinical phenotyping and prognosis.

- 1 Department of Computer Science, University of California Davis, Davis, CA.
- 2 Department of Internal Medicine, Division of Pulmonary, Critical Care, and Sleep Medicine, University of California Davis, Sacramento, CA.
- 3 Department of Respiratory Care, University of California Davis Health, Sacramento, CA.
- 4 Department of Emergency Medicine, University of Utah, Salt Lake City, UT.
- 5 Division of Informatics, Department of Public Health Sciences, University of California Davis, Davis, CA.
- 6 Department of Electrical and Computer Engineering, University of California Davis, Davis, CA.

The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. The funders had no role in study design, analysis, or interpretation, writing of the article, or the decision to submit for publication.

Drs. Rehm, Cortés-Puch, Chuah, and Adams conceived and designed the study. Dr. Rehm, Dr. Cortés-Puch, Dr. Kuhn, Mr. Nguyen, Dr. Fazio, Dr. Johnson, Dr. Anderson, and Dr. Adams coordinated and performed data collection and curation. All authors participated in analysis and interpretation of study data. Drs. Rehm, Cortés-Puch, and Adams wrote the first draft of the article and all authors participated in revision of the article for important intellectual content. Drs. Rehm, Anderson, Johnson, Chuah, and Adams obtained funding for the study, and Drs. Chuah and Adams provided overall study supervision. Drs. Chuah and Adams had full access to the data used in the study and take full responsibility for the integrity of the data and the accuracy of the data analysis. Dr. Adams was responsible for the final decision to submit.

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/ccxjournal>).

Drs. Rehm, Fazio, and Adams received funding from the National Heart, Lung, and Blood Institute related to the submitted work during

the study K12HL108964 (to Dr. Adams), F31HL144028 (to Dr. Rehm), and T32HL007013 (to Dr. Fazio). Drs. Johnson, Chuah, and Adams received funding from the UC Davis Collaborative for Diagnostic Innovation to support aspects of this work. Drs. Rehm, Chuah, and Adams had full access to all the data and Dr. Adams had final responsibility for the decision to submit. Drs. Rehm, Johnson, and Adams are listed on a patent application (U.S. patent application 62/420,943) related to the current work, and Drs. Johnson and Adams are cofounders of Certus Critical Care, a company working to translate aspects of the current work to the bedside. The remaining authors have disclosed that they do not have any potential conflicts of interest.

Address requests for reprints to: Jason Y. Adams, MD, MS, Division of Pulmonary, Critical Care, and Sleep Medicine, University of California Davis, 4150 V Street, Suite 3400, Sacramento, CA 95817. E-mail: jyadams@ucdavis.edu

REFERENCES

1. Bellani G, Laffey JG, Pham T, et al; LUNG SAFE Investigators; ESICM Trials Group: Epidemiology, patterns of care, and mortality for patients with acute respiratory distress syndrome in intensive care units in 50 countries. *JAMA* 2016; 315:788–800
2. Matthay MA, Zemans RL, Zimmerman GA, et al: Acute respiratory distress syndrome. *Nat Rev Dis Primers* 2019; 5:18
3. Needham DM, Yang T, Dinglas VD, et al: Timing of low tidal volume ventilation and intensive care unit mortality in acute respiratory distress syndrome. A prospective cohort study. *Am J Respir Crit Care Med* 2015; 191:177–185
4. Weiss CH, Baker DW, Weiner S, et al: Low tidal volume ventilation use in acute respiratory distress syndrome. *Crit Care Med* 2016; 44:1515–1522
5. Force ADT, Ranieri VM, Rubenfeld GD, et al: Acute respiratory distress syndrome: The Berlin definition. *JAMA* 2012; 307:2526–2533
6. Sjoding MW, Hofer TP, Co I, et al: Interobserver reliability of the Berlin ARDS definition and strategies to improve the reliability of ARDS diagnosis. *Chest* 2018; 153:361–367
7. Chen W, Janz DR, Shaver CM, et al: Clinical characteristics and outcomes are similar in ARDS diagnosed by oxygen saturation/FIO₂ ratio compared with PaO₂/FIO₂ ratio. *Chest* 2015; 148:1477–1483
8. Angus DC, Deutschman CS, Hall JB, et al: Choosing wisely® in critical care: Maximizing value in the intensive care unit. *Crit Care Med* 2014; 42:2437–2438
9. Martínez-Balzano CD, Oliveira P, O'Rourke M, et al; Critical Care Operations Committee of the UMass Memorial Healthcare Center: An educational intervention optimizes the use of arterial blood gas determinations across ICUs from different specialties: A quality-improvement study. *Chest* 2017; 151:579–585
10. Herasevich V, Yilmaz M, Khan H, et al: Validation of an electronic surveillance system for acute lung injury. *Intensive Care Med* 2009; 35:1018–1023

11. Azzam HC, Khalsa SS, Urbani R, et al: Validation study of an automated electronic acute lung injury screening tool. *J Am Med Inform Assoc* 2009; 16:503–508
12. Wayne MT, Valley TS, Cooke CR, et al: Electronic “Sniffer” systems to identify the acute respiratory distress syndrome. *Ann Am Thorac Soc* 2019; 16:488–495
13. McKown AC, Brown RM, Ware LB, et al: External validity of electronic sniffers for automated recognition of acute respiratory distress syndrome. *J Intensive Care Med* 2019; 34:946–954
14. Yetisgen-Yildiz M, Bejan CA, Wurfel MM: Identification of patients with acute lung injury from free-text chest x-ray reports. In: Proceedings of the 2013 Workshop on Biomedical Natural Language Processing. Sofia, Bulgaria, Association for Computational Linguistic, 2013, pp10–17
15. Gutierrez G, Ballarino GJ, Turkan H, et al: Automatic detection of patient-ventilator asynchrony by spectral analysis of airway flow. *Crit Care* 2011; 15:R167
16. Blanch L, Villagra A, Sales B, et al: Asynchronies during mechanical ventilation are associated with mortality. *Intensive Care Med* 2015; 41:633–641
17. Beitler JR, Sands SA, Loring SH, et al: Quantifying unintended exposure to high tidal volumes from breath stacking dyssynchrony in ARDS: The BREATHE criteria. *Intensive Care Med* 2016; 42:1427–1436
18. Adams JY, Lieng MK, Kuhn BT, et al: Development and validation of a multi-algorithm analytic platform to detect off-target mechanical ventilation. *Sci Rep* 2017; 7:14980
19. Sottile PD, Albers D, Higgins C, et al: The association between ventilator dyssynchrony, delivered tidal volume, and sedation using a novel automated ventilator dyssynchrony detection algorithm. *Crit Care Med* 2018; 46:e151–e157
20. Rehm GB, Han J, Kuhn BT, et al: Creation of a robust and generalizable machine learning classifier for patient ventilator asynchrony. *Methods Inf Med* 2018; 57:208–219
21. Lucangelo U, Bernabé F, Blanch L: Respiratory mechanics derived from signals in the ventilator circuit. *Respir Care* 2005; 50:55–65
22. van Drunen EJ, Chiew YS, Chase JG, et al: Expiratory model-based method to monitor ARDS disease state. *Biomed Eng Online* 2013; 12:57
23. Sundaresan A, Chase JG, Shaw GM, et al: Model-based optimal PEEP in mechanically ventilated ARDS patients in the intensive care unit. *Biomed Eng Online* 2011; 10:64
24. López V, Fernández A, García S, et al: An insight into classification with imbalanced data: Empirical results and current trends on using data intrinsic characteristics. *Information Sci* 2013; 250:113–141
25. Rehm GB, Kuhn BT, Delplanque JP, et al: Development of a research-oriented system for collecting mechanical ventilator waveform data. *J Am Med Inform Assoc* 2018; 25:295–299
26. Pedregosa F, Varoquaux G, Gramfort A, et al: Scikit-learn: Machine learning in python. *J Machine Learning Res* 2011; 12:2825–2830
27. Breiman L: Random forests. *Machine Learning* 2001; 45:5–32
28. Lanspa MJ, Gong MN, Schoenfeld DA, et al; The National Heart, Lung, and Blood Institute Prevention and Early Treatment of Acute Lung Injury (PETAL) Clinical Trials Network: Prospective assessment of the feasibility of a trial of low-tidal volume ventilation for patients with acute respiratory failure. *Ann Am Thorac Soc* 2019; 16:356–362
29. Bellani G, Pham T, Laffey JG: Missed or delayed diagnosis of ARDS: A common and serious problem. *Intensive Care Med* 2020; 46:1180–1183
30. Michael W, Sjoding RCH: Recognition and appropriate treatment of the acute respiratory distress syndrome remains unacceptably low. *Critical care medicine* 2016; 44:1611–1612
31. Herasevich V, Tsapenko M, Kojacic M, et al: Limiting ventilator-induced lung injury through individual electronic medical record surveillance. *Crit Care Med* 2011; 39:34–39
32. Solti I, Cooke CR, Xia F, et al: Automated classification of radiology reports for acute lung injury: Comparison of keyword and machine learning based natural language processing approaches. *IEEE* 2009; 2009:314–319
33. Afshar M, Joyce C, Oakey A, et al: A computable phenotype for acute respiratory distress syndrome using natural language processing and machine learning. *AMIA Annu Symp Proc* 2018; 2018:157–165
34. Fan-Minogue H, Maslove D, Lamb P, et al: Extracting computational and semantic features from portable chest X-rays for diagnosis of acute respiratory distress syndrome. *AMIA Jt Summits Transl Sci Proc* 2013; 2013:64
35. Chbat NW, Chu W, Ghosh M, et al: Clinical knowledge-based inference model for early detection of acute lung injury. *Ann Biomed Eng* 2012; 40:1131–1141
36. Reamaroon N, Sjoding MW, Lin K, et al: Accounting for label uncertainty in machine learning for detection of acute respiratory distress syndrome. *IEEE J Biomed Health Inform* 2019; 23:407–415
37. Rivello ED, Kiviri W, Twagirumugabe T, et al: Hospital incidence and outcomes of the acute respiratory distress syndrome using the Kigali modification of the Berlin definition. *Am J Respir Crit Care Med* 2016; 193:52–59
38. Coravos A, Khozin S, Mandl KD: Developing and adopting safe and effective digital biomarkers to improve patient outcomes. *NPJ Digit Med* 2019; 2:631–635
39. Galloway CD, Valys AV, Shreibati JB, et al: Development and validation of a deep-learning model to screen for hyperkalemia from the electrocardiogram. *JAMA Cardiol* 2019; 4:428–436
40. Hannun AY, Rajpurkar P, Haghpanahi M, et al: Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network. *Nat Med* 2019; 25:65–69
41. Tjepkema-Cloostermans MC, da Silva Lourenço C, Ruijter BJ, et al: Outcome prediction in postanoxic coma with deep learning. *Crit Care Med* 2019; 47:1424–1432
42. Calfee CS, Delucchi K, Parsons PE, et al; NHLBI ARDS Network: Subphenotypes in acute respiratory distress syndrome: Latent class analysis of data from two randomised controlled trials. *Lancet Respir Med* 2014; 2:611–620

43. Madotto F, Pham T, Bellani G, et al; LUNG SAFE Investigators and the ESICM Trials Group: Resolved versus confirmed ARDS after 24 h: Insights from the LUNG SAFE study. *Intensive Care Med* 2018; 44:564–577
44. Rubenfeld GD, Caldwell E, Granton J, et al: Interobserver variability in applying a radiographic definition for ARDS. *Chest* 1999; 116:1347–1353
45. Wainberg M, Merico D, DeLong A, et al: Deep learning in biomedicine. *Nat Biotechnol* 2018; 36:829–838
46. Thille AW, Esteban A, Fernández-Segoviano P, et al: Comparison of the Berlin definition for acute respiratory distress syndrome with autopsy. *Am J Respir Crit Care Med* 2013; 187:761–767
47. Zeiberg D, Prahlad T, Nallamothu BK, et al: Machine learning for patient risk stratification for acute respiratory distress syndrome. *PLoS One* 2019; 14:e0214465
48. Calfee CS, Janz DR, Bernard GR, et al: Distinct molecular phenotypes of direct vs indirect ARDS in single-center and multi-center studies. *Chest* 2015; 147:1539–1548