Supervised Machine Learning for the Early Prediction of Acute Respiratory Distress Syndrome (ARDS)

Sidney Le, BA[1]; Emily Pellegrini, MEng[1]; Abigail Green-Saxena, PhD*[1]; Charlotte Summers, BM, PhD[2]; Jana Hoffman, PhD[1]; Jacob Calvert, MSc[1]; Ritankar Das, MSc[1]

[1] Dascena Inc., Oakland, CA, United States

[2] Department of Medicine, University of Cambridge School of Clinical Medicine, Cambridge, United Kingdom

*Corresponding author:
Dr. Abigail Green-Saxena
Dascena, Inc.
414 13th Street
Suite #500
Oakland, CA 94612
abigail@dascena.com

ABSTRACT

Purpose: Acute respiratory distress syndrome (ARDS) is a serious respiratory condition with high mortality and associated morbidity. The objective of this study is to develop and evaluate a novel application of gradient boosted tree models trained on patient health record data for the early prediction of ARDS.

Materials and Methods: 9919 patient encounters were retrospectively analyzed from the Medical Information Mart for Intensive Care III (MIMIC-III) data base. XGBoost gradient boosted tree models for early ARDS prediction were created using routinely collected clinical variables and numerical representations of radiology reports as inputs. XGBoost models were iteratively trained and validated using 10-fold cross validation.

Results: On a hold-out test set, algorithm classifiers attained area under the receiver operating characteristic curve (AUROC) values of 0.905 when tested for the detection of ARDS at onset and 0.827, 0.810, and 0.790 for the prediction of ARDS at12-, 24-, and 48-hour windows prior to onset, respectively.

Conclusion: Supervised machine learning predictions may help predict patients with ARDS up to 48 hours prior to onset.

Key Words: Acute respiratory distress syndrome; intensive care unit; machine learning; clinical decision support systems; electronic health records; medical informatics

Abbreviations: AECC, American-European Consensus Conference; APACHE, Acute Physiologic Assessment and Chronic Health Evaluation; ARDS, Acute Respiratory Distress Syndrome; AUROC, Area Under Receiver Operating Characteristic; CDS, Clinical Decision Support; DOR, Diagnostic Odds Ratio; ED, Emergency Department; EHR, Electronic Health Records; GCS, Glasgow Coma Scale; ICU, Intensive Care Unit; ICD-9, International Statistical Classification of Diseases version 9; INR, International Normalised Ratio; LIPS, Lung Injury Prediction Score; LR+, Positive Likelihood Ratio; LR-, Negative Likelihood Ratio; MAP, Mean Arterial Pressure; MIMIC III, Medical Information Mart for Intensive Care version III; MLA, Machine Learning Algorithm; PEEP, Positive End Expiratory Pressure; P/F ratio, PaO2/FiO2 ratio; PP, Pulse Pressure; ROC, Receiver Operating Characteristic; SAPS, Simplified Acute Physiology Score; SOFA, Sequential Organ Failure Assessment; WBC, White Blood Cell Count

INTRODUCTION

Acute respiratory distress syndrome (ARDS) is a clinical syndrome characterized by hypoxemia in the presence of non-cardiogenic pulmonary edema, and is associated with severe inflammation.[1] ARDS is estimated to affect at least 190,000 patients per year in the United States[2] and has been cited as one of the leading causes of admission to the intensive care unit (ICU)[3,4], with mortality rates ranging between 30%-55%.[5] The wide variation in reported incidence[6] and mortality rates[2,7-14] may relate to difficulties in the recognition and diagnosis of ARDS. Despite high mortality rates and high rates of ICU utilization associated with ARDS, it is still critically misdiagnosed and underdiagnosed in intensive care units on a global scale.[1,5,15]

Difficulty in accurately diagnosing ARDS may be explained by a number of factors. These include differences in risk factors and etiologies, the availability of diagnostic tools, the quality and interpretation of chest radiographs, and general clinician ability to recognize ARDS.[7,16] The inability of healthcare providers to process the volume of clinical data generated while caring for critically ill patients has been cited as another potential reason for poor ARDS recognition.[17,18] The most recent Berlin definition[19] of ARDS was developed in 2012 in response to issues regarding the reliability and validity of the 1994 American-European Consensus Conference (AECC) definition.[20] Although the Berlin definition has addressed many of the limitations of the AECC definition,[19-21] identifying ARDS in diverse clinical settings remains dependent on some subjectivity of the diagnosing clinician.[22] Clinicians' ability to separate ARDS from other heterogeneous causes of respiratory failure is limited,[21,23,24] and it can often be difficult to diagnose ARDS in patients who have underlying medical problems with similar symptoms.[25]

Because ARDS treatment options have limited efficacy, there is an interest in identifying patients most at risk of developing ARDS for early prevention strategies, such as antiplatelet therapy.[7,26-28] Early identification of such patients could also improve treatment options by enabling early clinical trial enrollment.[29,30] The opportunity for preventing ARDS onset is constrained to a narrow window, with onset a median 2 days after hospital admission.[7,28,31,40] Despite advances in our understanding of ARDS pathogenesis, no biomarker has been shown to reliably predict ARDS.[32-34,41,42] Therefore, developing clinical decision support (CDS) methods to assist clinicians in the accurate and early prediction of ARDS is a valuable approach to improve patient monitoring, diagnosis, treatment, and outcomes.

CDS technologies have the ability to differentiate between groups of patients with similar conditions, and are useful in informing treatment decisions and improving patient outcomes.[33,35] They have recently been proposed as a method to improve early ARDS detection.[17,36,37] Through informed data analysis, CDS models can analyze relevant patient data from large electronic health record (EHR) databases and identify cohorts of patients with similar disease progression. We hypothesize that supervised machine learning can be used to improve ARDS detection and early ARDS prediction prior to onset. Here, we describe the development and analysis of a novel application of supervised machine learning model CDS for the detection and early prediction of ARDS. The benefit of such an approach is that when the model is implemented in clinical settings, healthcare providers can potentially identify patients at risk of developing ARDS before they deteriorate, thus facilitating effective resource allocation and identifying those patients most likely to benefit from increased monitoring and care.

MATERIALS AND METHODS

Data selection

Data were obtained from the Medical Information Mart for Intensive Care III (MIMIC-III) database, which consists of the inpatient ICU encounters at Beth Israel Deaconess Medical Center between 2001 and 2012.[38] The MIMIC-III publication states that, "the project was approved by the Institutional Review Boards of Beth Israel Deaconess Medical Center (Boston, MA) and the Massachusetts Institute of Technology (Cambridge, MA). Requirement for individual patient consent was waived because the project did not impact clinical care and all protected health information was deidentified."[38] To ensure consistent encoding of data, only data collected with the MetaVision clinical information system were used. All patient data collected using MetaVision was from patients admitted during or after 2008.

We applied additional inclusion criteria (**Figure 1**) to focus the scope of our study. Only patients with age data available and at least 18 years of age were included. Patient stays that did not have at least one observation of each required measurement type (see below) were excluded. Finally, we included only patient stays that had durations within a specified window. The upper limit on length of stay was set at 1000 hours (approximately 41.7 days), in order to account for outliers and transcription errors. The lower limit was dependent on lookahead, and the final study population sizes are listed in **Table 1**. For example, to predict for up to 48 hours before onset of ARDS using a five-hour window, 53 hours of patient data are required for inclusion.

We note that, in order to simulate the use case as a screening tool for the general population, the patient population under consideration was not restricted to mechanically ventilated patients, unlike other ARDS studies such as Taoum et al.[39] and Neto et al.[40] We have also analyzed separately a subpopulation in which patients are required to have experienced at least one hour of mechanical ventilation to be included in the study population (**Supplemental Table 1**).

Table 1. Number of encounters included in analysis.

Requirement				
All MIMIC-III encounters	53432			
Age exists, age at least 18	53332			
Metavision	23593			
At least 1 observation of each required measurement	22752			
Offset (hours)	**0**	**12**	**24**	**48**
Qualifying stay duration (duration ≥ offset + 5 hours)	21728	20388	15527	9251

Data extraction

Beginning at the first recorded measurement, raw measurements entered into the EHR for each patient stay were binned into one-hour intervals and averaged or summed within bins to produce a single, summarizing value per interval. Antibiotics, urine output, dobutamine, dopamine, epinephrine, norepinephrine, and phenylephrine measurements were summed, and all other clinical measurements listed in **Supplemental Table 2** were averaged. Encoding the data in this way transformed the measurements into discrete time series with consistent time steps, which were more readily handled by the algorithm. Not all raw measurements were available at all hours, so missing values were filled using last-one carry forward imputation. This is a natural imputation method for clinical measurements; observations of a raw measurement are expected to be dependent on the previous observations.[41,42]

For each patient stay, we took the vector of measurements using a five hour window. Where appropriate, we also concatenated the differences in measurement values between time steps. In this way, at prediction time, a supervised machine learning technique such as gradient boosted tree ensembles is able to access trend information and covariance structure with respect to time windows. This procedure of transforming time series problems into supervised learning problems has been used in our previous work.[43]

Models were developed using quantitative clinical features taken from the patient EHR, but patients were only required to have age, heart rate, respiratory rate, temperature, diastolic and systolic blood pressure, and SpO2 available. Other quantitative clinical features were included if available, and replaced with "missing" values where not available. Quantitative clinical features included for analysis in patient subpopulations with no required mechanical ventilation, and with at least one hour of mechanical ventilation, are listed in **Supplemental Table 2**. The organ dysfunction feature listed in **Supplemental Table 2** is defined to be the number of the following criteria which are met at a given time: systolic blood pressure < 90 mmHg; lactate > 2.0 mmol/L; platelet count < 100000 µL; and international normalized ratio > 1.5. The machine learning algorithm which we applied in this study is capable of learning from the distribution of missing values and can still gain information from relatively sparse features. We included only those patient stays which contained at least one measurement of each of the required features.

We extracted radiology reports and preprocessed them for use in our algorithm. Radiology reports are not expected to be present for every patient stay; moreover, it is valuable information if a patient stay does not have any radiology reports generated. Radiology reports contain complex information concerning clinician insight and the health of a patient. If the reports were present, our experimental design was able to access that complex information for machine learning, and if the reports were not present, the MLA was able to learn information about the lack thereof. Using the Doc2Vec text encoding scheme,[44] radiology reports were converted into numerical feature vectors. The Doc2Vec encoding network uses the relationships between words and their neighbors, as well as the relationship between paragraphs within a text, to generate a numerical embedding. These embeddings are crucial features in our experimental design for similar reasons; they allow the machine learning algorithm to access text representations of the clinical reality. These numerical embeddings are able to retain much of the relational structure of the text as a feature vector, without necessarily having to retain information about the literal text. The Doc2Vec encoding network, as implemented in the Python package gensim,[45] was trained on tokenized training texts, preprocessed to remove numbers and non-alphanumeric characters. This corpus of training texts was composed of 117,902 radiology reports, drawn from our training data. Once all training texts were observed and network weights updated, training procedures were frozen, and we then used the fully-trained Doc2Vec encoding network to infer the feature vectors for all of the radiology reports, similarly tokenized and preprocessed. These feature vectors were concatenated onto the existing quantitative clinical variables for patient stays where radiology reports were available. For patient stays where radiology reports were not available, vectors of the same size, containing missing values, were concatenated to the existing variables.

Gold standard and definition of onset time

In order to generate gold standard labels for ARDS, we followed the Berlin definition[19] as operationalized in Neto et al.[40] By examining the patient data for the co-occurrence of positive end expiratory pressure (PEEP) above or equal to 5 cmH_2O and PaO_2/FiO_2 ratio (P/F ratio) below or equal to 300 mmHg, we encoded positive class labels as 1 and negative class labels as 0. The mention of bilateral opacities or infiltrates in the patient's radiology report was also required for a positive class label. In order to ensure the acute nature of ARDS onset, we did not consider as ARDS positive any encounter involving a tracheostomy procedure within the first 72 hours of their ICU admission. The onset time for ARDS was set as the time of first co-occurrence for the PEEP and P/F ratio criteria, and prediction time was set to some number of hours prior to this onset. Thus a model described as a 24-hour model is a model for predicting ARDS 24 hours prior to onset by this co-occurrence definition.

We note that this gold standard does not determine the extent to which respiratory failure can be attributed to cardiac failure or fluid overload, which is a departure from the Berlin definition; we elaborate this limitation in the Discussion. We also emphasize that the measurements used to determine this gold standard were not used in development or training of the machine learning algorithm used in this study. In pilot experiments, we were able to verify the implementation of ARDS used in this study reproduced ARDS incidence rates observed in Neto et al.[40]

Experimental methods

For the purposes of evaluation, we reserved 10% of the patient stays within the MIMIC-III dataset, chosen at random, as a hold-out dataset and used only the remaining 90% to train, validate, and iterate our predictive models. This hold-out data represented unseen new data and could be used to gauge performance of machine learning algorithms in the setting of novel data prediction. Although we were primarily interested in prediction at 24-hours prior to onset, we also trained models for detection of ARDS at onset and prediction of ARDS at 12-hours and 48-hours prior to onset.

All predictive models described in this paper were instances of the XGBoost gradient boosted tree model,[46] implemented using the Python package. XGBoost is a state-of-the-art tree ensemble method that builds progressively on the loss generated by weak decision tree base learners. XGBoost is capable of learning quickly and effectively from large amounts of data, and is flexible to the point that it is able to learn even from missing data. By making use of this capability, we were able to construct predictive models that do not require radiographs or radiology reports to make meaningful predictions. It is important to note that decision tree models, including tree ensembles, do not make distributional assumptions, and so are well-suited for settings where specifying a generative distribution is difficult.

Three of the available hyperparameters for XGBoost were selected using exhaustive grid search five-fold cross validation, performed exclusively on the training data. Five folds for hyperparameter tuning is the default for hyperparameter grid search due to considerations of computational constraints, as implemented in Scikit-learn.[47] The hyperparameters tuned were number of base learners, the learning rate, and the maximum depth of a base learner. The hyperparameters were tuned across ranges of values centered around 1000, 0.1, and 5 for number of base learners, learning rate, and maximum depth, respectively. The values selected as the centers were determined by iteratively narrowing the grid search range. These three hyperparameters affected the values the internal model parameters took over the course of training, and thus also significantly contributed to the final model parameters.

The XGBoost predictive models were all iteratively trained and tested using ten-fold cross validation with early stopping mechanisms in order to prevent overfitting. In this validation paradigm, the data were partitioned into ten random segments, or folds. Training occurred on nine of the folds, and the remaining fold was used to monitor performance for overfitting. Each of the ten models trained were then tested on the hold-out test set partitioned prior to hyperparameter tuning, and the final metrics reported were averages for the metrics across the ten models. Metrics reported include area under the receiving operator curve (AUROC), standard deviation for the area metrics, sensitivity, specificity, accuracy, recall, diagnostic odds ratio (DOR), and positive and negative likelihood ratios (LR+ and LR-, respectively). Once all models were trained, we evaluated the performance of the models in predicting the ARDS labels of the hold-out set, and the same performance metrics were reported.

In addition to this main set of experiments validating the effectiveness of our algorithm as a screening tool developed and tested on the general patient population, we conducted an additional analysis in which we developed and evaluated the same algorithm using only mechanically ventilated patients. All procedures, from partitioning into training and hold-out test sets to hyperparameter tuning and training, were performed identically in this additional experiment.

RESULTS

Among stays meeting the inclusion criteria of **Table 1** and with qualifying duration for the 0-, 12-, 24-, and 48-hour offsets, respectively 296 (1.362%), 179 (0.877%), 107 (0.689%), and 25 (0.270%) of stays were labeled as ARDS positive according to the gold standard. Analogously, for the stays meeting the inclusion criteria of **Supplemental Table 1** and with qualifying duration for the 0-, 12-, 24-, and 48-hour offsets, respectively 288 (3.199%), 174 (1.998%), 104 (1.366%), and 25 (0.455%) of stays were labeled as ARDS positive. Demographic data of all patient encounters from the MIMIC-III dataset are presented in **Table 2,** including the distribution of admissions to various wards in the ICU and the distribution of physiologic derangement, represented by MEWS scores, at admission.

Using the training data, we performed five-fold cross validation on every combination of hyperparameter values in our pre-specified hyperparameter ranges. In total there were 72 different hyperparameter combinations, and with five-fold cross validation, a total of 360 models were fit on the training data. The evaluation metric used to determine the best performing hyperparameter combination was AUROC. The hyperparameters selected to train the final models were: 1000 base learners, a learning rate of 0.03, and a base learner maximum depth of six partition levels.

ARDS onset detection and prediction performance is summarized by the Receiver Operating Characteristic (ROC) curves in **Figure 2.** ROC curves show sensitivity (the fraction of ARDS positive cases that received an ARDS positive label) as a function of 1−specificity (the fraction of ARDS negative cases that received an ARDS positive label). Operating points of approximately 0.80 sensitivity were selected for each model to facilitate comparisons of performance. Each ROC curve represents the average performance under 10-fold cross validation. The classifier demonstrated an AUROC of 0.905, 0.827, 0.810, and 0.790 for early ARDS detection and prediction on the test set at 0, 12, 24, and 48 hours prior to onset, respectively (**Figure 2**). AUROC curves demonstrated high sensitivity and specificity of algorithm predictions for ARDS onset up to 48 hours in advance on the test set.

Table 2. Demographics of subjects included in analyses. Percentage values may not add to 100 due to rounding. Demographics calculated for patients with stay durations of at least 48 hours.

	Characteristic	**Value (%)**

Gender	Female	4432 (44.7)
	Male	5487 (55.3)
Age (years)	18-29	356 (3.6)
	30-39	394 (4.0)
	40-49	902 (9.1)
	50-59	1691 (17.1)
	60-69	2335 (23.6)
	70+	4241 (42.8)
ICU Ward Admission	ICU	3937 (39.7)
	SICU	1939 (19.5)
	CSRU	1515 (15.3)
	CCU	1303 (13.1)
	TSICU	1225 (12.4)
MEWS Severity at Admission	0	4099 (41.3)
	1	1093 (11.0)
	2	1301 (13.1)
	3	1193 (12.0)
	4	996 (10.0)
	5	641 (6.5)
	6	340 (3.4)
	7	165 (1.7)
	8	54 (0.5)
	9	22 (0.2)
	10	11 (0.1)
	11	2 (0.0)
	12	2 (0.0)
Median length-of-stay (IQR) days	4 (2, 7)	

Multiple performance metrics are shown in **Table 3**, including AUROC,, sensitivity and specificity, representing a variety of clinically relevant assessments for the general patient population. All metrics were calculated with common operating points near sensitivity = 0.80 in order to allow for direct comparisons. Testing performance metrics for the patient population with at least one hour of mechanical ventilation are reported in **Supplemental Table 3**, and the AUROC curves associated with the performance are shown in **Supplemental Figure 1.** In this mechanically ventilated population, our classifier demonstrated AUROC performance of 0.843, 0.858, 0.810, and 0.790 for early ARDS detection and prediction on the test set at 0-, 12-, 24-, and 48- hours prior to onset, respectively.

Table 3. Model performance metrics on the training and testing (hold-out) sets at 0-, 12-, 24-, and 48-hour prediction windows. AUROC: area under the receiving operator curve; DOR: diagnostic odds ratio; LR+ and LR-: positive and negative likelihood ratios, respectively. Values presented are means and standard deviations for the metrics across 10 folds.

	Onset	**12-hour**	**24-hour**	**48-hour**
AUROC	0.905 (0.009)	0.827 (0.015)	0.810 (0.035)	0.790 (0.079)
Sensitivity	0.806 (0.000)	0.789 (0.000)	0.818 (0.000)	0.667 (0.000)
Specificity	0.823 (0.014)	0.828 (0.052)	0.683 (0.073)	0.852 (0.063)
F1	0.109 (0.006)	0.079 (0.015)	0.020 (0.018)	0.015 (0.004)
DOR	19.477 (1.829)	19.704 (5.953)	10.452 (3.664)	13.175 (4.485)
LR+	4.576 (0.354)	4.938 (1.253)	2.719 (0.666)	5.058 (1.495)
LR-	0.235 (0.004)	0.255 (0.017)	0.269 (0.028)	0.393 (0.032)
Accuracy	0.825 (0.010)	0.839 (0.045)	0.817 (0.160)	0.851 (0.061)
Recall	0.774 (0.000)	0.732 (0.017)	0.427 (0.369)	0.333 (0.000)

As shown in **Supplemental Table 4**, antibiotics administration appears to yield a significant amount of information about the classifier across all prediction times in the general patient population. However, there are few other observable trends in feature importances that are consistent. It should be noted that the stochastic nature of the XGBoost algorithm, which extends to the subset of columns which it considers in individual trees in its ensemble, limits the interpretability of feature importances.

DISCUSSION

We have described a method for the early prediction of ARDS using supervised machine learning models. Model classifiers attained AUROC values of 0.827, 0.810, and 0.790 for the prediction of ARDS at 12-, 24, and 48- hours prior to onset, respectively (**Figure 2**). In addition to high AUROC values, model classifiers demonstrated high performance for the detection and prediction of ARDS in regards to sensitivity, specificity, F1, DOR, L+, L-, accuracy, and recall (**Table 3**). We developed these models using quantitative clinical features extracted from the patient EHR data, as well as numerical representations of radiology reports. Our approach circumvents the issues associated with keyword-based text analysis by using higher-level representations of the text in radiology reports. These numerical representations are used as features in our model, alongside the patient quantitative, structured data. The use of this structured data to complement radiographic reports mitigates delays in obtaining chest radiograph information. In these ways, the method we describe diversifies and improves upon existing approaches for the prediction of ARDS.

Inability to anticipate which patients are likely to develop ARDS is a major obstacle to early intervention and to prevention studies.[48] Epidemiologic data suggest that the syndrome is rarely present at the time of hospital admission or initial emergency department (ED) evaluation, but develops over a period of hours to days in subsets of at-risk patients.[49-53] Therefore, evaluating model performance at >24 hours preceding onset is valuable because it facilitates identification of patients who would benefit from targeted ARDS interventions. Alerting systems for the long horizon prediction of ARDS have been validated in similar studies of mechanically ventilated patients and those with moderate hypoxia.[54,55]

While rule-based systems have been used to screen patients for ARDS by analyzing patient EHR data,[48,56-59] non rules-based CDS systems are capable of efficiently incorporating complex patient data sets and are therefore less reliant on clinician subjectivity. Several studies have focused on the development of non rules-based ARDS detection systems[17,54,60-64] and represent a promising means for addressing the issues clinicians face when identifying pre-existing ARDS. However, our study was designed to address the need for CDS systems that can predict ARDS onset sufficiently in advance to provide clinicians with time to undertake preventative measures. A previous study by Taoum et al. described a novel approach for early prediction of ARDS using continuous physiological signals of heart rate, respiratory rate, peripheral arterial oxygen saturation and mean airway blood pressure.[39] Results indicated that ARDS can be detected in the early phases of occurrence with a sensitivity of 65% and a specificity of 100%, on average 39 hours prior to onset.[39] However, this study was undertaken on a small dataset, which limits generalizability andr elies on minute-by-minute samples from physiological monitors to detect ARDS, which introduces a potential barrier to hospital integration and which ignores the benefits of unstructured clinical notation data.[39] In contrast, our method uses relatively sparsely sampled structured data, such as vital signs and lab tests, in addition to unstructured notation data. The method of Zaglam et al. requires chest radiographs to be obtained before it may assess the presence or absence of ARDS, which may hinder the early diagnosis or prediction of ARDS, and does not use text data.[60] While the rules-based method of Herasevich et al. uses unstructured radiographic report data, it does so by searching reports for a list of keywords, which is vulnerable to misdiagnosis arising from the presence of keywords mentioned in ruling out diagnosis, and which potentially neglects more complicated textual indications of ARDS.[56] In contrast, our use of Doc2Vec enables the extraction of rich, contextual information from unstructured texts, including information concerning chest radiographs highly relevant to ARDS. Our approach does so without explicitly requiring radiographs to generate a prediction score, which allows the tool to be used as a screening tool for the general population.

Our supervised machine learning models demonstrate high diagnostic metrics for ARDS recognition and prediction in general patient populations (**Table 3**). The testing curves of **Figure 2** demonstrate the model's strength in diagnosis at the time of ARDS onset, with an AUROC value 0.905 for the general patient population. These metrics outperform those reported in other studies.[62] While the quality of diagnostic metrics decay as they are made increasingly early prior to ARDS onset, the 12-hour prediction of ARDS offers operating points with high sensitivity and specificity. **Table 3** illustrates a clinically relevant operating point with sensitivity of 0.806 and specificity of 0.823. Early prediction of ARDS onset offers opportunities for increased patient monitoring, possible prevention,[26-28] and the development of novel preventative measures.

In the mechanically ventilated subpopulation, our supervised machine learning models demonstrated a similarly high level of diagnostic performance for ARDS recognition and prediction (**Supplemental Table 3**). Models in both the mechanically ventilated subpopulation and the general population achieved high sensitivity and specificity for 12-hour prediction of ARDS, with an operating point with sensitivity of 0.778 and specificity of 0.810 in the mechanically ventilated subpopulation. At the time of ARDS onset, an AUROC of 0.843 was observed in this subpopulation, compared to AUROC of 0.905 in the general population. The performance 12 hours prior to onset was higher in the mechanically ventilated subpopulation, with an AUROC of 0.858 compared to AUROC of 0.827. Overall we observed similar performance in both patient populations.

We emphasize several limitations of our study. First, the results of our study may not generalize to analogous experiments conducted with a definition of ARDS other than the Berlin definition, or different implementations thereof. Indeed, we use the implementation of the Berlin definition used by Neto et al. [40], which does not assess the extent to which respiratory failure can be attributed to cardiac failure or fluid overload. Strictly speaking, this is a departure from the Berlin definition but, by our assessment, it would be difficult to unambiguously implement this criterion using available data and without introducing bias, for the following reasons. Determining if respiratory failure can be fully attributed to hydrostatic lung edema, in the absence of a risk factor, requires an objective assessment. However, it is not always clear which assessment to undertake, which complicates the incorporation of such assessments into a gold standard [65]. Moreover, it has been reported that 30% of ARDS cases include a component of hydrostatic lung edema [65], so erroneous adjudication of these cases could substantially under-label encounters as ARDS positive or otherwise introduce significant bias into ARDS labeling. Second, this study was a retrospective analysis, which may not translate to prospective improvements in clinical settings. In particular, the retrospective performance metrics we report cannot capture the complex interaction of clinicians with the information such a tool would provide, or the limitations of ARDS prevention and treatment options. Finally, this study concerned a single-center study of ICU data and therefore the results may not translate to other clinical settings or wards, especially wards of less intensive care. In future work we hope to develop and evaluate this tool in a variety of live clinical settings.

CONCLUSION

This analysis demonstrates the use of a gradient boosted tree model for the early prediction and identification of ARDS using retrospective patient data. The algorithm developed in this study may assist both in recruitment for ARDS clinical trials and the improved prediction and early recognition of ARDS.

ACKNOWLEDGEMENTS

We gratefully acknowledge Anna Lynn-Palevsky and Christine Giang for their work in editing this manuscript.

Author contributions: RD had full access to all of the data in the study and takes responsibility for the integrity of the data and the accuracy of the data analysis. Conceptualization, SL, JC, JH, CS, RD; Methodology, SL, JC, RD; Formal Analysis, SL, JC; Investigation, SL, JC; Writing--Original Draft, RD, JH, JC, SL, AGS, CS, and EP; Supervision, RD and JH.

Conflict of interest statement: SL, EP, AGS, JH, JC, and RD are employees of Dascena.

Funding/Support: This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

REFERENCES

1. Matthay MA, Ware LB, Zimmerman GA. The acute respiratory distress syndrome. *J Clin Invest.* 2012; 122:2731-2740.
2. Rubenfeld GD, Caldwell E, Peabody E, et al. Incidence and outcomes of acute lung injury. *N Engl J Med.* 2005;353:1685–1693.
3. Dziadzko MA, Herasevich V, Pickering BW. Predicting outcomes from respiratory distress: does another score help to solve the problem? *Crit Care Med.* 2016; 44:1437-1438.
4. Barrett ML, Smith MW, Elixhauser A, et al: Utilization of Intensive Care Services, 2011. HCUP Statistical Brief #185. Rockville, MD, Agency for Healthcare Research and Quality, 2014
5. Bellani G, Laffey JG, Pham T, et al. 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.
6. Summers C, Singh NR, Worpole L, et al. Incidence and recognition of acute respiratory distress syndrome in a UK intensive care unit. *Thorax.* 2016; 71:1050-1.
7. Confalonieri M, Salton F, Fabiano F. Acute respiratory distress syndrome. *Eur Respir Rev.* 2017; 26:1-7.
8. Brun-Buisson C, Minelli C, Bertolini G, et al. Epidemiology and outcome of acute lung injury in European intensive care units. Results from the ALIVE study. *Inten Care Med.* 2004; 30:51–61.
9. Villar J, Blanco J, Añón JM, et al. The ALIEN study: incidence and outcome of acute respiratory distress syndrome in the era of lung protective ventilation. *Inten Care Med.* 2011; 37:1932–1941.

10. Caser EB, Zandonade E, Pereira E, et al. Impact of distinct definitions of acute lung injury on its incidence and outcomes in Brazilian ICUs: prospective evaluation of 7,133 patients. *Crit Care Med.* 2014; 42:574–582.
11. Cochi SE, Kempker JA, Annangi S, et al. Mortality trends of acute respiratory distress syndrome in the United States from 1999–2013. *Ann Am Thorac Soc.* 2016; 13:1742–1751.
12. Erickson SE, Martin GS, Davis JL, et al. Recent trends in acute lung injury mortality: 1996–2005. *Crit Care Med.* 2009; 37:1574–1579.
13. Sigurdsson MI, Sigvaldason K, Gunnarsson TS, et al. Acute respiratory distress syndrome: nationwide changes in incidence, treatment and mortality over 23 years. *Acta Anaesthesiol Scand.* 2013; 57:37–45.
14. Spieth PM, Guldner A, Gama de Abreu M. Acute respiratory distress syndrome: basic principles and treatment. *Anaesthesist.* 2017; 66:539-552.
15. Fröhlich S, Murphy N, Doolan A, *et al*. Acute respiratory distress syndrome: underrecognition by clinicians. J Crit Care. 2013;28:663-8.
16. Sjoding MW, Cooke CR, Iwashyna TJ, et al. Acute respiratory distress syndrome measurement error. Potential effect on clinical study results. *Ann Am Thorac Soc.* 2016; 13:1123–1128.
17. Reamaroon N, Sjoding MW, Lin K, Iwashyna TJ, Najarian K. Accounting for label uncertainty in machine learning for detection of acute respiratory distress syndrome. *IEEE J Biomed Health Inform.* 2019; 23407-415.
18. Clark BJ, Moss M. The acute respiratory distress syndrome: Dialing in the evidence? *JAMA.* 2016; 315:759-761.
19. ARDS Definition Task Force, Ranieri VM, Rubenfeld GD, et al. Acute respiratory distress syndrome: the Berlin definition. JAMA. 2012; 307:2526–2533.
20. Bernard GR, Artigas A, Brigham KL, et al. The American-European Consensus Conference on ARDS: definitions, mechanisms, relevant outcomes, and clinical trial coordination. *Am J Respir Crit Care Med.* 1994; 149:818-824.
21. Bauman ZM, Gassner MY, Coughlin MA, et al. Lung injury prediction score is useful in predicting acute respiratory distress syndrome and mortality in surgical critical care patients. *Crit Care Res Pract.* 2015;157408.
22. Coudroy R, Frat JP, Boissier F, *et al*. Early identification of acute respiratory distress syndrome in the absence of positive pressure ventilation: implications for revision of the Berlin criteria for acute respiratory distress syndrome. *Crit Care Med.* 2018;46:540-6.

23. Luhr OR, Karlsson M, Thorsteinsson A, et al. The impact of respiratory variables on mortality in non-ARDS and ARDS patients requiring mechanical ventilation. *Inten Care Med*. 2000; 26:508-517.

24. Rubenfeld GD, Caldwell E, Granton J, et al. Interobserver variability in applying a radiographic definition for ARDS. *Chest*. 1999; 116:1347–1353.

25. Tobin M, Manthous C. What is Acute Respiratory Distress Syndrome? *Am J Respir Crit Care Med*. 2017;196(9):16-17.

26. Erlich JM, Talmor DS, Cartin-Ceba R, et al. Prehospitalization antiplatelet therapy is associated with a reduced incidence of acute lung injury: a population-based cohort study. *Chest*. 2011; 139:289–295.

27. Chen W, Janz DR, Bastarache JA, et al. Prehospital aspirin use is associated with reduced risk of acute respiratory distress syndrome in critically ill patients: a propensity-adjusted analysis. *Crit Care Med*. 2015; 43:801–807.

28. Kor DJ, Carter RE, Park PK, et al. Effect of aspirin on development of ARDS in at-risk patients presenting to the emergency department: the LIPS-A randomized clinical trial. *JAMA*. 2016; 315:2406–2414.

29. Yadav H, Thompson BT, Gajic O. Fifty years of research in ARDS. Is Acute Respiratory Distress Syndrome a Preventable Disease? *Am J Respir Crit Care Med*. 2017; 195:725-736.

30. Festic E, Kor DJ, Gajic O. Prevention of acute respiratory distress syndrome. *Curr Opin Crit Care*. 2015; 21:82-90.

31. Gajic O, Malinchoc M, Comfere TB, et al. The Stability and Workload Index for Transfer score predicts unplanned intensive care unit patient readmission: Initial development and validation. *Crit Care Med*. 2008; 36:676–682.

32. Blondonnet R, Constantin JM, Sapin V, et al. A pathophysiologic approach to biomarkers in acute respiratory distress syndrome. *Dis Markers.* 2016; 3501373.

33. Ware LB, Calfee CS. Biomarkers of ARDS: what's new? *Inten Care Med*. 2016; 42:797-799.

34. Peer D. Precision medicine–delivering the goods? *Cancer Lett*. 2014; 352:2-3.

35. Janz DR, Ware LB. Biomarkers of ALI/ARDS: pathogenesis, discovery, and relevance to clinical trials. *Semin Respir Crit Care Med*. 2013; 34:537–548.

36. Sjoding MW, Hyzy RC. Recognition and appropriate treatment of the acute respiratory distress syndrome remains unacceptably low. *Critical Care Med.* 2016; 44:1611-1612.

37. Sjoding MW. Translating evidence into practice in acute respiratory distress syndrome: Teamwork, clinical decision support, and behavioral economic interventions. *Curr Opin. Crit Care*. 2017; 23:406-411.

38. Johnson AEW, Pollard TJ, Shen L, *et al.* MIMIC-III, a freely accessible critical care database. *Scientific Data*. 2016. DOI: 10.1038/sdata.2016.35. Available from: http://www.nature.com/articles/sdata201635

39. Taoum A, Mourad-chehade F, Amoud H. Early-warning of ARDS using novelty detection and data fusion. *Comp Biol Med*. 2018; 102:191-199.

40. Neto AS, Deliberato RO, Johnson AEW, *et al.* Mechanical power of ventilation is associated with mortality in critically ill patients: an analysis of patients in two observational cohorts. *Intens Care Med.* 2018; 44(11):1914-22.

41. Shao J, Zhong B. Last observation carry-forward and last observation analysis. *Stat Med*. 2003; 22:2429-41.

42. Ali MW, Talukder E. Analysis of longitudinal binary data with missing data due to dropouts. *J Biopharm Stat*. 2005; 15:993-1007.

43. Mohamadlou H, Lynn-Palevsky A, Barton C, et al. Prediction of acute kidney injury with a machine learning algorithm using electronic health record data. *Can J Kidney Health Dis*. 2018; 5:2054358118776326.

44. Le Q, Mikolov T. Distributed representations of sentences and documents. In: International conference on machine learning 2014 Jan 27 (pp. 1188-1196).

45. Gensim: Topic Modelling for Humans. Machine Learning Consulting, 10 Apr 2019. https://radimrehurek.com/gensim/about.html.

46. Chen T, Guestrin C. XGBoost: A Scalable Tree Boosting System. 22nd SIGKDD Conference on Knowledge Discovery and Data Mining. 2016.

47. Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: Machine Learning in Python. JMLR 12 2011; 2825-2830.

48. Gajic O, Dabbagh O, Park PK, et al. Early identification of patients at risk of acute lung injury: evaluation of lung injury prediction score in a multicenter cohort study. *Am J Respir Crit Care Med.* 2011; 183:462–470.

49. Fernandez-Perez ER, Yilmaz M, Jenad H, Daniels CE, Ryu JH, Hubmayr RD, Gajic O. Ventilator settings and outcome of respiratory failure in chronic interstitial lung disease. *Chest.* 2008;133:1113–1119.

50. Fowler AA, Hamman RF, Good JT, Benson KN, Baird M, Eberle DJ, Petty TL, Hyers TM. Adult respiratory distress syndrome: risk with common predispositions. *Ann Intern Med.* 1983;98:593–597.

51. Gong MN, Thompson BT, Williams P, Pothier L, Boyce PD, Christiani DC. Clinical predictors of and mortality in acute respiratory distress syndrome: potential role of red cell transfusion. *Crit Care Med.* 2005;33:1191–1198.

52. Hudson LD, Milberg JA, Anardi D, Maunder RJ. Clinical risks for development of the acute respiratory distress syndrome. *Am J Respir Crit Care Med.* 1995;151:293–301.

53. Pepe PE, Potkin RT, Reus DH, Hudson LD, Carrico CJ. Clinical predictors of the adult respiratory distress syndrome. *Am J Surg.* 1982;144:124–130.

54. Zeiberg D, Prahlad T, Nallamothu BK, *et al.* Machine learning for patient risk stratification for acute respiratory distress syndrome. *PloS One.* 2019; 14:e0214465.

55. Azzam HC, Khalsa SS, Urbani R, Shah CV, Christie JD, Lanken PN, Fuchs BD. Validation study of an automated electronic acute lung injury screening tool. *JAMIA*. 2009; 16:503-508.

56. Herasevich V, Yilmaz M, Khan H, et al. Validation of an electronic surveillance system for acute lung injury. *Inten Care Med*. 2009; 35:1018-1023.

57. Herasevich V, Tsapenko M, Kojicic M, *et al.* Limiting ventilator-induced lung injury through individual electronic medical record surveillance. *Crit Care Med.* 2011; 39:34-9.

58. Lin CY, Kao KC, Tian YC, et al. Outcome scoring systems for acute respiratory distress syndrome. *Shock.* 2010; 34:352–357.

59. Trillo Alvarez C, Cartin-Ceba R, Kor DJ, et al. Acute lung injury prediction score: derivation and validation in a population based sample. *Eur Respir J*. 2001; 37:604-609.

60. Zaglam N, Jouvet P, Flechelles O, et al. Computer-aided diagnosis system for the Acute Respiratory Distress Syndrome from chest radiographs. *Computers Biol Med.* 2014; 52:41-48.

61. Soto GJ, Kor DJ, Park PK, et al. Lung injury prediction score in hospitalized patients at risk of acute respiratory distress syndrome. *Crit Care Med*. 2016; 44:2182-2191.

62. Schenck EJ, Oromendia C, Torres LK, Berlin DA, Choi AMK, Siempos II. Rapidly Improving ARDS in Therapeutic Randomized Controlled Trials. *Chest*. 2018; pii: S0012-3692: 32582-0.

63. McKown AC, Brown RM, Ware LB, *et al.* External validity of electronic sniffers for automated recognition of acute respiratory distress syndrome. *J Intens Care Med.* 2017; 0885066617720159.

64. Wayne MT, Valley TS, Cooke CR, *et al.* Electronic “Sniffer” Systems to Identify the Acute Respiratory Distress Syndrome. *Annals Am Thorac Society*. 2019; 16:488-95.

65. Redant S, Devriendt J, Botta I, Attou R, De Bels D, Honoré PM, Pierrakos C. Diagnosing acute respiratory distress syndrome with the Berlin definition: Which technical investigations should be the best to confirm it?. Journal of translational internal medicine. 2019 Mar 29;7(1):1-2.

FIGURE CAPTIONS

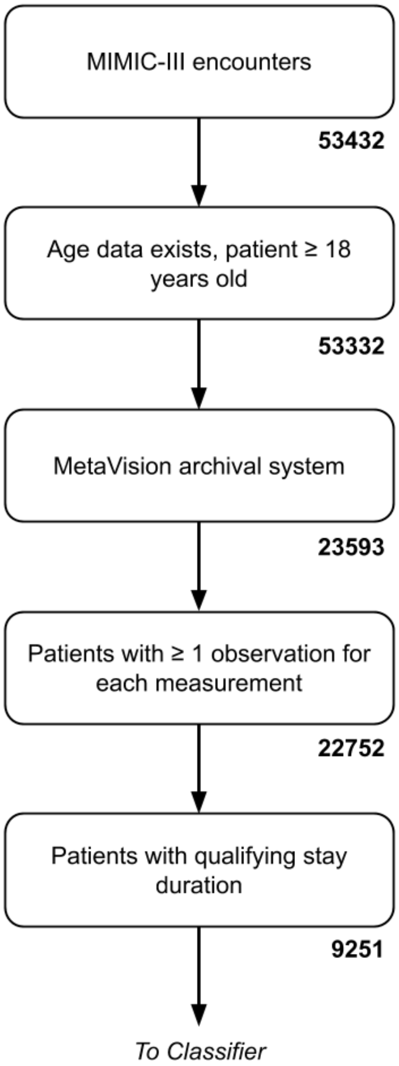

MIMIC-III encounters
53432
Age data exists, patient ≥ 18 years old
53332
MetaVision archival system
23593
Patients with ≥ 1 observation for each measurement
22752
Patients with qualifying stay duration
9251
To Classifier

Figure 1. Inclusion criteria for patient encounters in the MIMIC-III dataset. The final inclusion criteria is dependent on prediction lookahead; the value presented here reflects the 48-hour prediction, which filters most stringently.

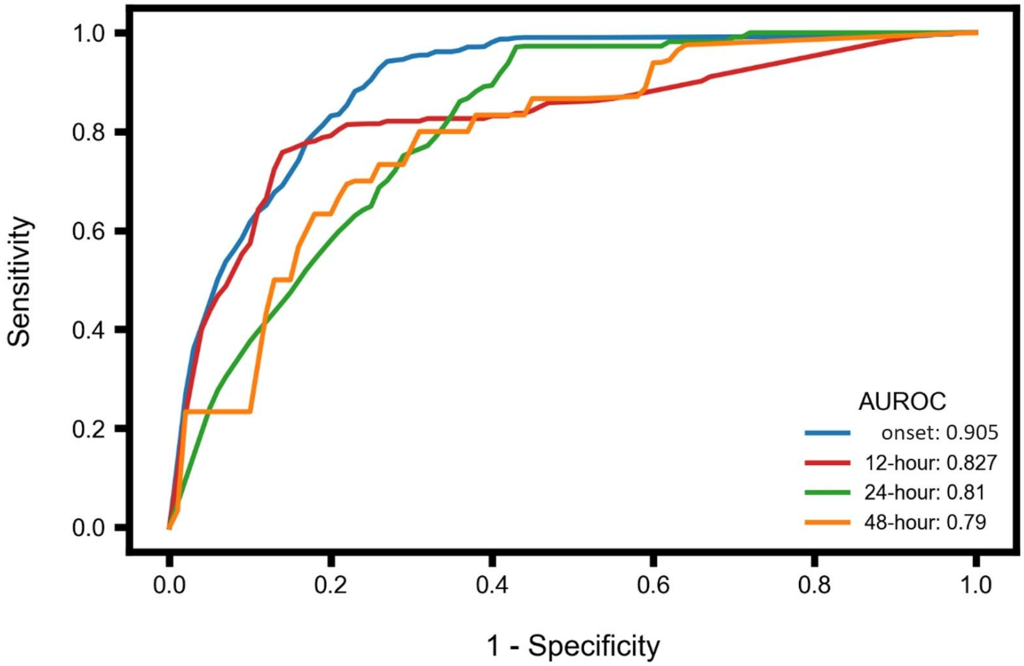


Figure 2. Area Under Receiver Operating Characteristic (AUROC) curves and values for ARDS onset detection and prediction at 12, 24, and 48 hours prior to onset. AUROC performance of XGBoost models on a separate hold-out test set for early ARDS prediction, up to 48 hours prior to onset. Curves are averaged across 10 folds.

SUPPLEMENTAL MATERIALS

Supplemental Table 1. Inclusion table of patient subpopulation for analysis with at least one hour of mechanical ventilation.

Requirement				
All MIMIC-III encounters	53432			
Age exists, age at least 18	53332			
Metavision	23593			
At least 1 observation of each required measurement	22752			
Mechanical ventilation	9133			
Offset (hours)	**0**	**12**	**24**	**48**
Qualifying stay duration (duration ≥ offset + 5 hours)	9001	8706	7609	5483

Supplemental Table 2. Clinical features included for analysis in patient subpopulations with no required mechanical ventilation, and with at least one hour of mechanical ventilation. GCS = Glasgow Coma Scale; HR = Heart Rate; INR = International Normalised Ratio; MAP = Mean Arterial Pressure; PP = Pulse Pressure; SpO2 = Peripheral Capillary Oxygen Saturation; WBC = White Blood Cell Count

Clinical Features
Age
Antibiotics
Bilirubin
Blood Culture
Creatinine
Diastolic BP
Fluid Bolus
GCS
HR
INR
Lactate
MAP
Organ Dysfunction
PP
Platelets
Resp. Rate
SpO2
Systolic BP
Temp.
Urine Output
WBC
pH

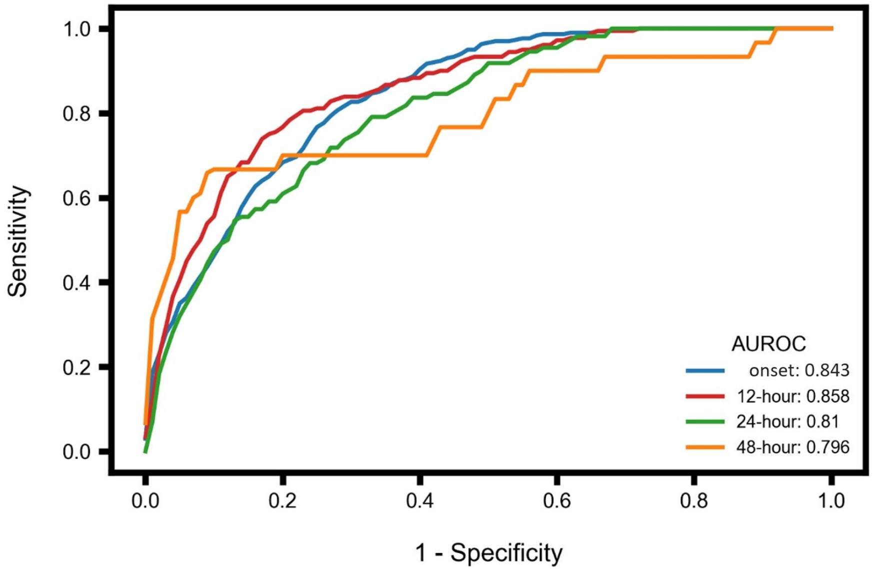
Supplemental Table 3. Model performance metrics at 0-, 12-, 24-, and 48-hour detection and prediction windows, on the patient subpopulation test set with at least one hour of mechanical ventilation. AUROC: area under the receiving operator curve; DOR: diagnostic odds ratio; LR+ and LR-: positive and negative likelihood ratios, respectively. Values presented are means and standard deviations for the metrics across 10 folds.

	Onset	12	24	48
AUROC	0.843 (0.015)	0.858 (0.022)	0.810 (0.044)	0.796 (0.112)
Sensitivity	0.800 (0.000)	0.778 (0.000)	0.818 (0.000)	0.667 (0.000)
Specificity	0.733 (0.040)	0.810 (0.072)	0.671 (0.115)	0.910 (0.127)
F1	0.159 (0.019)	0.142 (0.040)	0.066 (0.029)	0.094 (0.098)
DOR	11.274 (2.169)	17.107 (6.838)	11.382 (7.924)	89.235 (117.989)
LR+	3.055 (0.434)	4.579 (1.520)	2.888 (1.441)	30.412 (39.330)
LR-	0.274 (0.016)	0.276 (0.027)	0.279 (0.049)	0.375 (0.073)
Accuracy	0.734 (0.039)	0.808 (0.070)	0.671 (0.113)	0.913 (0.129)
Recall	0.767 (0.000)	0.722 (0.000)	0.727 (0.000)	0.300 (0.105)

Supplemental Table 4. Feature importances. Most important features across prediction times on the patient subpopulation test set with no required mechanical ventilation, determined by the relative average information gain across all trees in the ensemble and scaled according to lookahead time. GCS = Glasgow Coma Scale; HR = Heart Rate; INR = International Normalised Ratio; MAP = Mean Arterial Pressure; PP = Pulse Pressure; SpO2 = Peripheral Capillary Oxygen Saturation; WBC = White Blood Cell Count

	Onset	12	24	48
Age	0.004115	0	0	0.053510
Antibiotics	0.261102	1	1	0.743631
Bilirubin	0.001402	0.008645	0	0.583681
Blood cultures	0	0	0	0
Creatinine	0.056243	0.00803	0	0.079805
Diastolic BP	0.04635	0.015093	0.022645	0.155653
Fluid Bolus	0.030919	0.009978	0.042103	0.103617
GCS	1	0.355869	0	0.418873
HR	0.018309	0.028341	0.117448	0.552735
INR	0.001856	0	0	0.316773
Lactate	0.045153	0.012959	0	0.417665
MAP	0.10657	0.022658	0	0.101288
Organ Dysfunction	0.001482	0.002276	0	0.603887
PP	0.024534	0.018626	0.068241	0.245032
Platelets	0.003704	0.004354	0.019733	0.327886
Resp. Rate	0.051007	0.056339	0.020247	0.413433
SpO2	0.053115	0.065467	0	1
Systolic BP	0.022387	0.039432	0.017186	0.162508
Temp.	0.005873	0.014421	0.02879	0.143157
WBC	0.016341	0.010967	0	0.221133
Urine output	0	0	0	0.003511
pH	0.10487	0.011362	0.040173	0.232714

Due to the data extraction process, features are represented in the MLA by multiple columns in the data matrices, each of which are treated as individual latent features. The importance scores presented here are the average importance across those latent features, scaled within lookahead times to give a relative average importance score.



Supplemental Figure 1. Area Under Receiver Operating Characteristic (AUROC) curves and values for ARDS onset detection and prediction at 12, 24, and 48 hours prior to onset, on the patient subpopulation test set with at least one hour of mechanical ventilation. Curves are averaged across 10 folds.