TITLE: A Data-Driven Approach to Patient Risk Stratification for Acute Respiratory Distress Syndrome (ARDS)

AUTHORS: Tejas Prahlad

INTRODUCTION

Acute Respiratory Distress Syndrome (ARDS) is a condition in which fluid and inflammatory cells permeate the lungs preventing effective gas transport. ARDS is linked to complications such as nosocomial infections and mortality. Over 70% of patients with ARDS are diagnosed late or not at all. In the US, ARDS is estimated to affect 34 per 100,000 patients per year, and studies show it has a 28-day mortality rate of 20-40%. These facts show the importance of making predictions as early and accurately as possible. [1] Previous studies, such as the LIPS study, have created models that identify patients who are at high risk for developing ARDS based on information available during the first six hours of a patient's hospital stay. We test the generalizability of LIPS in our study population and improve upon this model by leveraging the structured contents of the electronic health record (EHR) to identify patients who will develop ARDS during their hospital stay as early and accurately as possible.

METHODS:

Study Population:

In this retrospective cohort study, we considered all adult critically ill patients admitted to the University Hospital at the University of Michigan (UM) between January 2016 and March 2016. We included patients who received at least 3L of supplemental oxygen for at least 2 hours at some point during their first 6 days of hospital stay, or developed hypoxic respiratory failure.

Case Identification:

Ground truth labels were generated for a randomly selected subset of the population. Cases were identified based on chart review by clinicians. Utilizing the Berlin Criteria, up to 6 clinicians determined whether the patient developed ARDS and when. In addition, clinicians provided confidence in each evaluation on a scale of 1-4. Disagreements were resolved using a majority vote.

Data Extraction:

Patient data were extracted from longitudinal electronic health records from the UM Research Data Warehouse (RDW). We included both baseline characteristics and data recorded within the

[1] "Acute respiratory distress syndrome - The Lancet." http://www.thelancet.com/clinical/diseases/acute-respiratory-distress-syndrome. Accessed 28 Jun. 2017.

first six hours of the admission. Baseline characteristics pertain to data available at the time of admission, including socio-demographics (e.g., age, race, gender). Data recorded within the first six hours include: vital signs (e.g., temperature, heart rate etc.), in-hospital locations (e.g., ED, ICU etc.), laboratory values (e.g., pH level, WBC level, K level etc.), medication data (medication administration information) and other longitudinal data (e.g., respiratory support information, measures of neurologic function etc.). This is explained more succinctly in Table 2.

Statistical Methods:

Data Preprocessing. Patient data were continuous, discrete, or categorical. Through a series of preprocessing steps, we mapped all data to a vector of ones and zeros. Each variable may be measured only once (e.g., age) or multiple times (e.g., heart rate). We refer to these two variables types as time-invariant and time-varying, respectively.

We represented numerical variables (e.g., age) with up to 5 binary features. These features were determined in a data-driven fashion by dividing the data into quintiles (when possible). If it was impossible to split the variable of interest into 5 bins, (e.g., more than 20% of the data share the same value) we considered a smaller number of bins. Additional details are provided in APPENDIX A.

For numerical variables in which we had multiple measurements over time, (i.e., time-varying numerical variables) we had an additional step. We first summarized all observations for a given patient (within the first 6 hours of their hospital stay) using 6 statistics: minimum, maximum, mean, median, standard deviation and interquartile range. These summary statistics are then mapped to up to 5 bins based on quintiles (as described above).

Categorical data (e.g., medications) with d categories were mapped to a d-dimensional binary vector, in which each dimension (i.e., feature) corresponds to a different category. Laboratory tests values were split into 5 categories using the recorded values and provided reference ranges for "very high", "high", "normal", "low" and "very low." In addition, since often the absence of a test can be informative, we included a 6th category: "missing." For categorical variables with multiple measurements (e.g., multiple WBC) we simply turn the feature "on" (i.e., set it equal to 1) if the corresponding category is observed and "off" (i.e., set it equal to 0) otherwise. In terms of laboratory data, this means that a patient may have both high and low categories turned on, if both were observed during the observation period. For medication data, the final vectors indicate which medications were ever administered during the first 6 hours of the patient's hospital stay.

Learning Algorithm. We used regularized logistic regression to learn a mapping from the feature space (described above) and the ARDS labels. We used L_2 regularization to not unnecessarily eliminate potentially important features. Our learning pipeline is illustrated in Figure 1. We split the data into an 80-20 stratified train-test split. We learned the regression parameters on the training set. The cost hyperparameter (which trades of the complexity of the model and the performance on the training set) was selected using grid-search and 10-fold cross validation optimizing for the area under the receiver operating characteristic curve (AUROC). Performance was measured on the separate test set, again using the AUROC. This process was repeated 1000 times since the split was chosen randomly. To estimate generalization performance, we aggregated performance across all 1000 test sets and report performance in terms of the median AUROC and the empirical 95th confidence interval (i.e., the 2.5th and 97.5th percentiles). In all our analyses, predictions are only evaluated after patients have met the inclusion criteria stated above. This is done to ensure that test performance closely mimics generalization performance.

Apart from tracking the models' performances, the model weights were also recorded and the median values of weights across training sets was used to estimate feature importance. We record the features associated with the top ten positive weights (i.e., risk factors) and the top ten negative weights (i.e., protective factors). To measure statistical significance, we used a resampling method. I.e., we computed p-values based on the empirical distribution of each weight as follows:

$$p = 2 * \min(le, ge)$$

where, le represents the fraction of estimates for that coefficient that are less than or equal to 0 while ge represents the fraction of estimates for that coefficient that are greater than or equal to 0. When le or ge equaled zero we set the p=0.001, since we had 1000 bootstrapped samples. We considered a level of significance of 0.05.

In addition to this main contribution, we explore the use semi-supervised approaches to further improve predictive performance as well as time-series predictions to provide more useful, timely information to physicians. The semi-supervised methods aim to leverage the clear majority of unlabeled data during the training phase to further increase the model's ability to generalize. The methods explored so far include label propagation and self-training for classification. In terms of time-series predictions, the aim is to provide predictions as to whether a patient will develop ARDS in the next 24 or 48 hours at regular time intervals. These methods are discussed in more detail in Appendix B.

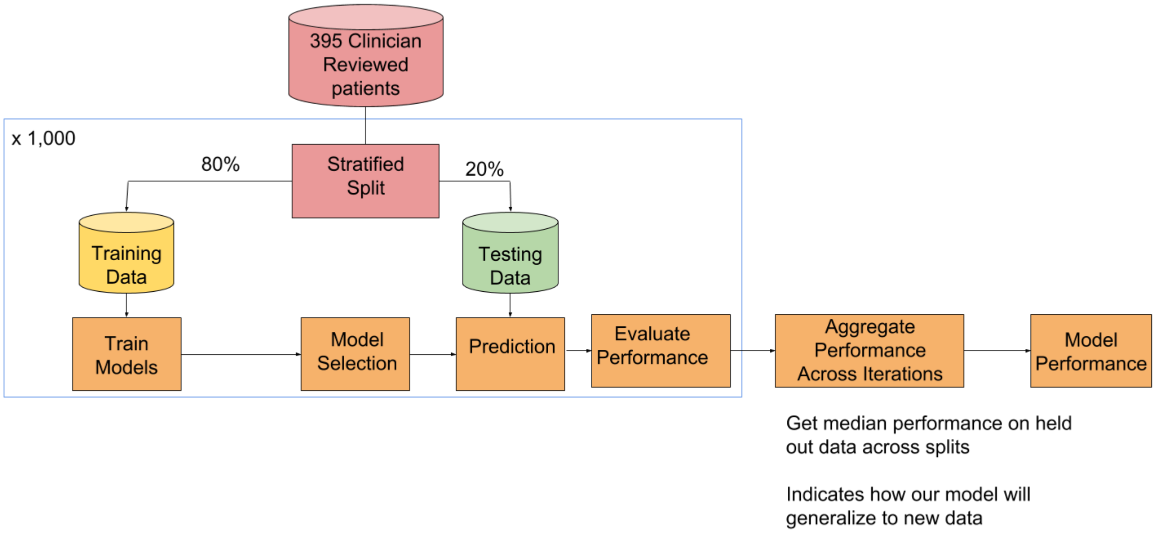
395 Clinician Reviewed patients
x 1,000
80%
Stratified Split
20%
Training Data
Testing Data
Train Models
Model Selection
Prediction
Evaluate Performance
Aggregate Performance Across Iterations
Model Performance
Get median performance on held out data across splits
Indicates how our model will generalize to new data

Figure 1: Overall training and testing pipeline

RESULTS:

After the inclusion criteria was applied to the UMHS data, the final study population included 3514 patients. Of this a subset of 395 patients were reviewed by clinicians. Selected characteristics of our study population are provided in Table 1.

Out of the 395 patients, 41 patients developed ARDS (positive cases) at some point during their first 6 days of hospital stay. Of the 395 reviewed cases, in 354 of them, all the physicians agreed (no conflicts). In 40 cases, there was only one physician who disagreed with the majority and 1 case where out of four physician reviews, two of them thought that the patient had developed ARDS while the other two did not.

In terms of predictive performance, the risk prediction model utilizing only EHR features demonstrated good discriminative performance AUROC = 0.81 (95% C.I. = 0.59-0.93). However, by leveraging the additional predictive utility of the LIPS variables we achieve a boost in performance AUROC = 0.83 (95% C.I. = 0.64-0.96), which is significantly better than the EHR model. Both models significantly outperform the LIPS model applied to our dataset (Table 3). The EHR+LIPS^{+} model identifies positive cases a median of 38.47 hours in advance of the earliest clinical diagnosis (Figure 2).

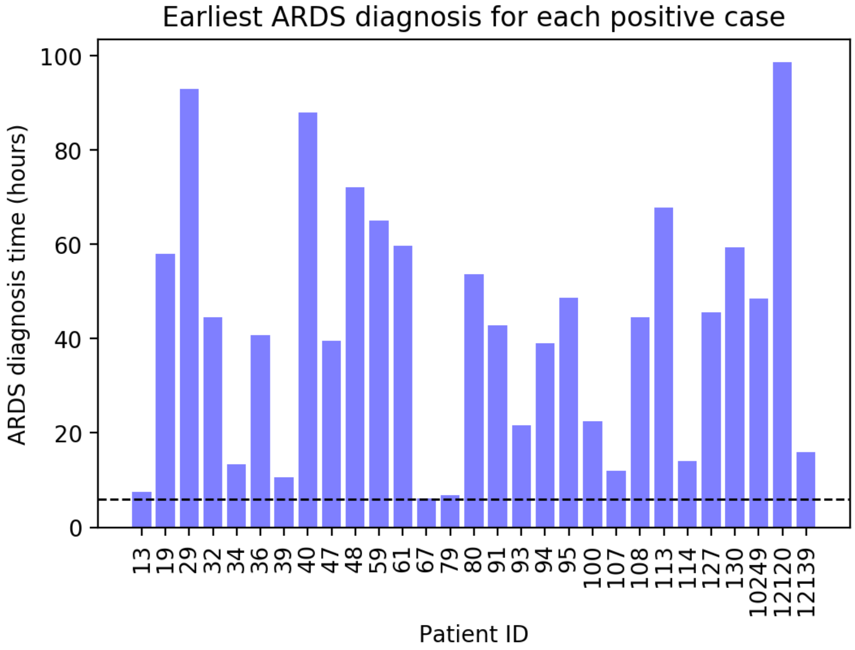

Figure 2: Earliest clinical diagnosis times and our model prediction time (dotted line)

The top ten risk factors (positive feature weights) and the top ten protective features (negative feature weights) are shown in Table 4. Within the top 10 risk factors, we find measures of neurologic function such as level of alertness and responsiveness recorded by the physician. Apart from these, well known risk factors such as high potassium levels, old age and whether the patient has undergone invasive surgery or not, are also present. With respect to the protective factors, low diastolic pressure is the highest ranked feature, followed by others like low heart rate and low respiratory rate.

Tables:

Table 1: Study Population Characteristics

Feature Category	Size
Patient Demographics (total)	**3514**
Patients reviewed	395
Positive cases (ARDS diagnosis)	41
Median age (IQR)	62.0 (51.0-71.0)
Female (%)	47.12
Race (%)	
Caucasian	85.14
African-American	8.94
Asian	1.37
Native Hawaiian or Pacific Islander	1.34
American Indian or Native Alaskan	0.37
Other	2.85
Clinical Characteristics	
Length of stay, d (median [IQR])	5.0 (2.5-6.0)
ARDS diagnosis time, h (median [IQR])	21.6 (4.6-48.6)
Median LIPS score (IQR)	3.0 (1.0-5.5)

Table 2: Breakdown of features included in the model

Feature Category	Example	Pre-processing	No. of Features	No. of Columns
Baseline Numerical	Age	Recursive binning	1	5
Baseline Categorical	Gender, Race	Binarized by category	3	12
Laboratory Test Flags	pH flag, WBC flag	Binarized by flag levels	21	126
Continuous Variables	Heart rate, platelet count	Recursive binning after summarization	32	616
Medication Data	Azithromycin, insulin	Binarized by administration	125	125
TOTAL			**182**	**884**

Table 3: Model Performance

Experiment	Description	Median AUROC (95% C.I.)
LIPS	LIPS scores	0.73 (0.53 – 0.88)
EHR	EHR features	0.81 (0.59 – 0.93)
EHR + LIPS^{+}	Both EHR and LIPS features	0.83 (0.64 – 0.96)

Table 4: Risk factors (top) and protective factors (bottom)

No.	Feature	Median Coefficient (95% C.I.)
1	Invasive Surgery (yes/no)	0.81 (0.020 – 4.810)
2	Sedated (yes/no)	0.51 (0.010 – 3.230)
3	Unresponsive (yes/no)	0.35 (0.010 – 3.400)
4	High Flow Nasal cannula (yes/no)	0.32 (0.010 – 2.180)
5	High Potassium Level	0.27 (0.010 – 2.590)
6	High Diastolic Blood Pressure	0.25 (0.002 – 1.860)
7	Age between 72-90	0.24 (-0.010 – 2.140)
8	Oriented (yes/no)	0.22 (0.003 – 1.940)
9	Non-Invasive Surgery (yes/no)	0.22 (0.004 – 1.410)
10	High Respiratory Rate	0.21 (0.007 – 1.590)
1	Low Diastolic Blood Pressure	-0.32 (-2.560 – -0.007)
2	Age between 65-72	-0.27 (-2.060 – -0.004)
3	Low Heart Rate	-0.25 (-1.680 – -0.010)
4	Low IV input	-0.23 (0.1510 – -0.010)
5	High P_{O2}/ Fi_{O2}	-0.23 (-1.710 – -0.003)
6	High Temperature	-0.23 (-1.730 – -0.003)
7	High Oxygen Saturation	-0.22 (-1.860 – -0.008)
8	Low Respiratory Rate	-0.21 (-1.360 – -0.010)
9	Low Fi_{O2}	-0.21 (-1.540 – -0.001)
10	Normal Total Bilirubin Level	-0.17 (-1.430 – -0.001)

REFERENCES:

APPENDIX A

Recursive Binning Strategy

This strategy was employed as a pre-processing step to handle all the continuous-valued variables in the dataset. The algorithm basically bins the values of a variable into a maximum of $n = 5$ bins (quintiles) and if the distribution of the values makes that impossible or counter-intuitive (e.g., if 40% of the values are the same or lie in one bin it would not make sense to split this bin into 2), then into some $n < 5$ bins. Quintiles were chosen since they have been proven to capture the distribution of the data better than other binning methods.

APPENDIX B

Semi-Supervised Learning Approaches

As the majority of the data we use is unlabeled (88.76%), it is important to leverage this vast amount of data so that we are confident that our model can generalize well. The first method that was explored was self-training for classification where a classifier trained on the labeled data, is applied to classify and consequently "label" the unlabeled data, which is again used for training. The algorithm is detailed below:

1. Learn the optimal classifier for the labeled training data alone
2. Using this classifier, classify and thus, label the unlabeled points in the dataset
3. Learn the optimal classifier using both the true labeled points as well as the newly labeled points
4. Evaluate the performance of this classification model using a held-out test set composed exclusively of truly labeled points

The other method explored was label propagation for classification. This approach aims to utilize clustering approaches to classify and "label" unlabeled points which can then be used for training the final logistic regression model. The algorithm is detailed below:

1. Using the label propagation method to spread the labels of truly labeled points to their neighbors, previously unlabeled points are now assigned labels
2. Now, using both truly labeled points and assigned points, a classifier is learned
3. 5-fold cross validation is performed to find the optimal hyper-parameters for both the label spreading and the classification algorithm

In terms of the label spreading algorithm the 2 hyper-parameters that we are interested in are the distance metric used (e.g. k-nearest neighbors, RBF kernel) and the neighborhood associated with the kernel (e.g. number of nearest neighbors and distance threshold). It is important to note that when cross-validation is performed, only truly labeled points are used to evaluate a particular model and not a point whose label was assigned to it by the algorithm. This is done to ensure that the model's true discriminative performance is measured and is not biased.

However, both distance metrics assume a uniform weighting on the features which is not optimal in our case. Hence a new distance metric is used which weights each of the feature components based on feature importance. This gives a better, more intuitive way of measuring distance between points so that the spreading of labels is more reliable.

Time-Series Predictions

Time series prediction basically involves providing probabilities with confidence intervals of the likelihood that a patient will develop ARDS within the next 24 hours, at regular time steps. In the experiments run, we use 2 hour intervals as this was advised by our clinical collaborators. The idea behind making these predictions builds on our single prediction method, which is then extrapolated using a sliding window approach. First, using just the first 6 hours of data, a model is learned to predict the outcome of a patient within the next 24 hours. This 6-hour data collection window is then moved forward 2 hours, and another prediction is made. This process continues until no data exists to verify our predictions or when the 24-hour prediction horizon hits the end of a patient stay. This has a much greater utility to clinicians than the previous single prediction method as this better captures the variability in the patient's time-varying signals. By providing doctors with constant predictions, it allows them to make decisions quicker and with more confidence.