Towards Reliable ARDS Clinical Decision Support: ARDS Patient Analytics with Free-text and Structured EMR Data

Emilia Apostolova, PhD¹, Amit Uppal, MD², Jessica E. Galarraga, MD, MPH³,⁴, Ioannis Koutroulis, MD, PhD⁵, Tim Tschampel⁶, Tony Wang PhD⁷, Tom Velez, PhD, JD⁶

¹Language.ai, Chicago, IL; ²NYU School of Medicine, Bellevue Hospital Center, New York, NY; ³MedStar Health Research Institute, Hyattsville, MD; ⁴MedStar Washington Hospital Center, Georgetown University School of Medicine, Washington, DC; ⁵Children’s National Health System, Washington, DC; ⁶Computer Technology Associates, Ridgecrest, CA; ⁷Imedacs, Ann Arbor, MI

Abstract

In this work, we utilize a combination of free-text and structured data to build Acute Respiratory Distress Syndrome (ARDS) prediction models and ARDS phenotype clusters. We derived ‘Patient Context Vectors’ representing patient-specific contextual ARDS risk factors, utilizing deep-learning techniques on ICD and free-text clinical notes data. The Patient Context Vectors were combined with structured data from the first 24 hours of admission, such as vital signs and lab results, to build an ARDS patient prediction model and an ARDS patient mortality prediction model achieving AUC of 90.16 and 81.01 respectively. The ability of Patient Context Vectors to summarize patients’ medical history and current conditions is also demonstrated by the automatic clustering of ARDS patients into clinically meaningful phenotypes based on comorbidities, patient history, and presenting conditions. To our knowledge, this is the first study to successfully combine free-text and structured data, without any manual patient risk factor curation, to build real-time ARDS prediction models.

Introduction

Critical care Clinical Decision Support (CDS) systems aim at the early identification and timely treatment of rapidly progressive, life-threatening conditions. In particular, ARDS (Acute Respiratory Distress Syndrome) is a significant cause of morbidity and mortality in the USA and worldwide¹,². Early recognition and evidence-based management of ARDS can limit the propagation of lung injury and significantly improve patient outcomes³.

To date, there exists no accurate and reliable way to anticipate which patients, presenting with respiratory distress, are likely to develop ARDS. Numerous prediction scores have been developed to assess ARDS prognosis and risk of death, such as Lung Injury Score (LIS)⁴, Lung Injury Prediction Score (LIPS)⁵, APPS (Age, Plateau, PaO2/FiO2 Score)⁶, Early Acute Lung Injury (EALI)⁷, and Modified ARDS Prediction Score (MAPS)⁸. Using a consensus process, a panel of experts convened in 2011 to develop the Berlin definition, focusing on addressing a number of limitations of prior definitions⁹. Still, the predictive validities of these tools and definitions have proven to be moderate, for example, as measured by area under receiver operating curve (AUC).

The difficulty in analyzing and predicting ARDS outcomes stems from the fact that this is both rare, and, at the same time, highly heterogeneous condition¹⁰. ARDS involves the interaction of multiple risk factors, past history, and current conditions, signs, and symptoms. Hospital alert systems typically rely on highly sensitive screening of structured data, such as vital signs and lab results, which, in the case of such rare conditions, are often associated with false clinical alarms resulting in “alarm fatigue”¹¹.

EMR data depends on what the clinician deems necessary to measure and record in the act of caring for the patient. EMR data is typically entered for the purposes of clinical documentation and billing¹², and thus not centered around the needs of real-time surveillance-based CDS systems. The combined physician care related variables and underlying patient-related contextual factors needed for a reliable ARDS risk evaluation are typically dispersed across the patient EMR record, and available at different times throughout the patient stay. Patient demographics, past medical and visit history, chronic conditions, risk factors, current signs and symptoms can be found in diverse combinations of structured elements and clinical notes (e.g. nursing notes, radiology reports, etc.), that record diagnosis and procedure codes, vital signs, lab orders and results, ventilation parameters, etc. The challenge for real-time surveillance-based CDS systems is accommodating for the variability and the availability of real-time electronic data and enabling accurate contextual interpretation of real-time patient data.

In this work, we utilize all available EMR patient information, in the form of structured data and free-text, for real-time predictive modeling. While our experiments are focused on identifying ARDS cases, the described method is
applicable to a variety of disease surveillance CDS use cases, needing information dispersed across the EMR patient record.

A second goal of this study is to utilize the combination of clinician knowledge and experience, and a data-driven approach to identify ARDS patients’ phenotypes and risk factors, acknowledging the need for targeted personalized treatments reflecting differences in treatment outcomes across patient subtypes.\textsuperscript{13–15}

**Dataset**

Clinical encounter data of adult patients were extracted from the MIMIC3 Intensive Care Unit (ICU) database.\textsuperscript{16} MIMIC3 consists of retrospective ICU encounter data of patients admitted into Beth Israel Deaconess Medical Center from 2001 to 2012. Included ICUs are medical, surgical, trauma-surgical, coronary, cardiac surgery recovery, and medical/surgical care units. MIMIC3 includes time series data recorded in the EMR during encounters (e.g. vital signs/diagnostic laboratory results, free text clinical notes, medications, procedures, etc.). The dataset contains data associated with over 58,000 ICU visits, including over 2 million free-text clinical notes and over 650,000 diagnosis codes.

For this study, in accordance with previous literature,\textsuperscript{17} we identified ARDS for adult patients older than 18 years with ICD-9 codes for severe acute respiratory failure and use of continuous invasive mechanical ventilation, excluding those with codes for acute asthma, COPD and CHF exacerbations.\textsuperscript{1} This resulted in 4,624 ARDS cases from a total of 48,399 adult ICU admissions. The ICU mortality rate in this population was approximately 59%, somewhat higher than expected for ARDS,\textsuperscript{18} suggesting that the algorithm used is capturing the most severe cases of ARDS, and thus introducing some level noise with possibly containing true ARDS cases marked as negative examples.

Our ARDS predictive model utilized data in the form of free-text clinical notes, ICD codes, and structured physiological and ventilator data. The structured data included in this analysis consists of anion gap (aniongap), albumin, bands, bicarbonate, bilirubin, creatine, chloride, glucose, hematocrit, hemoglobin, lactate, platelet, potassium, partial thromboplastin time (ptt), international normalized ratio (inr), prothrombin time (pt), sodium, bun, white blood cell count (wbc), heart rate (heartrate), systolic blood pressure (sysbd), diastolic blood pressure (diasbp), mean blood pressure (meanbp), respiratory rate (resperate), body temperature (tempc), peripheral capillary oxygen saturation (spo2), body mass index (bmi), gender, age, urine output (urine1). All variables are included as min, max, and mean values and are measured over the first 24 hours of ICU admission. The first 24 hour timeframe was chosen, as it has been reported that ARDS develops at a median of 30 hours after hospital admission.\textsuperscript{19} Thus, a 24-hour window provides for the gathering of enough structured data, while at the same time is early enough for real-time CDS.

**Descriptive Analytics**

ARDS patient characteristics and risk factors were first gathered with the help of experienced clinicians. Expert knowledge was gathered in the form of Concept Maps (Cmaps).\textsuperscript{20} Cmaps is a tool developed at the Institute for Human and Machine Cognition (IHMC) that enables collaborative knowledge creation, in the form of concepts, relations, and ontologies, with links to external resources and publications. A snippet of the developed ARDS Cmap developed by our clinical research team is shown in Figure 1.

For example, clinicians coded ARDS precipitant causes include sepsis, aspiration, traumatic injuries, burns, and drugs, including illicit drugs, such as cocaine, heroin, or prescription drugs, such as chemotherapeutic agents, etc.

Initially, the ARDS Cmap was used as a screening rule engine. In addition, the Cmap was used to identify risk factors that were later used in predictive models. Although rule engines based on Cmaps are evidence-based and tend to be highly sensitive, they tend to perform with subpar specificity (e.g. the Berlin definition achieved an AUC of 0.577). Such rule engines are also highly sensitive to missing data. In contrast, data-driven and Machine Learning algorithms have the potential to improve on rule engines from training on large datasets, learning from a variety of clinical data and response patterns, and are able to handle missing data. However, for machine learning to be effective in predicting highly heterogeneous conditions such as ARDS, training data requires both high precision labeling and the identification of features with adequate numbers of samples needed to separate classifications of ARDS classes and subclasses/phenotypes from non-ARDS patients.

\textsuperscript{1}Inclusion ICD9 Codes: 51881, 51882, 51884, 51851, 51852, 51853, 5184, 5187, 78552, 99592, 9670, 9671, 9672; Exclusion ICD9 Codes: 49391, 49392, 49322, 4280.
ICD Embeddings and Patient Vectors

We then looked for a data-driven approach to provide additional insight into ARDS patient characteristics, risk factors, and phenotypes.

Intuitively, even without any additional patient EMR data, clinicians viewing properly coded patient diagnosis codes (e.g. ICD codes in a problem list) are typically able to create a mental summary of the overall patient condition, including medical history, risk factors, presenting conditions. ICD codes are used to describe both current diagnoses (e.g. Pneumonia, unspecified organism; ICD9 486), but also a variety of additional patient information. For example, ICD codes can describe ARDS risk factors, such as patient’s history and chronic conditions (e.g. Chronic kidney disease; Personal history of malignant neoplasm; etc.); information regarding past and current treatments and procedures (e.g. Infection due to other bariatric procedure). In some cases, ICD codes contain information such as the patient age group and/or susceptibilities (e.g. Sepsis of newborn; Elderly multigravida); expected outcome (Encounter for palliative care); patient social history (e.g. Adult emotional/psychological abuse; Cocaine dependence); the reason for the visit, (e.g. Railway accidents; Motor Vehicle accidents).

Using ICD codes for statistical analysis and predictive models, however, poses a series of challenges. Patient ICD codes in EMRs tend to be sparse. There are numerous ICD codes (around 15,000 ICD9 codes and around 68,000 ICD10 codes), with only a very small subset of these applicable to a particular patient (e.g. MIMIC3 admissions have an average of 11 ICD codes). ICD codes also tend to co-occur and overlap. In addition, ICD coding can be, in some cases, subjective and dependent on numerous external factors.

However, the concurrence and mutual information of ICD codes over large data repositories can be utilized. For example, the fact that Pneumonia ICD codes are often accompanied with ICD codes describing Cough, Fever, Pleural effusion, etc. can be utilized to generate vector representations of ICD codes. Inspired by deep learning representation, such as word embeddings, it has been suggested that this medical code co-occurrence can be exploited to generate low-dimensional representations of ICD codes that may facilitate EMR data-based exploratory analysis and predictive modeling.

In this study, we utilized available MIMIC3 patient data to generate the ICD embeddings following the approach of Choi et al. In our approach, we generated a low-dimensional representation of the patient history, symptoms, risk factors, diagnosis, etc, by averaging the patient ICD code embeddings. We refer to this representation of the patient's medical history and clinical condition as Patient Context Vectors.

To generate ARDS patients groups sharing similar characteristics and risk factors, we then clustered the ARDS Patient Context Vectors via k-means clustering. Clinical review of the generated ARDS clusters determined the optimal number of clusters to be 10.
The Patient Context Vectors were able to clearly separate ARDS patient risk factors and conditions into clinically valid categories, such as Malignancy or Chronic Hepatic Disease. Figure 2 shows the frequency of patients in various clusters, sorted left-to-right according to mortality rate. Table 1 lists the 15 most representative ICD code descriptions (cluster centroids) for the 10 ARDS clusters. The cluster description was provided by clinicians reviewing the corresponding cluster centroids.

![Figure 2: Frequency and mortality rate of MIMIC3 ARDS patient clusters formed by clustering of averaged ICD embeddings.](image)

Interestingly, the manually clinician-curated Cmap appears to overlap to a large extent with the automatically derived ARDS patient clusters. For example, clinicians listed chemotherapeutic agents as a distinct risk factor for ARDS. Our data-driven analysis showed a distinct cluster of ARDS patients with ICD codes describing malignancy. Further analysis could delineate if this is a direct relationship between chemotherapeutic agents and the development of ARDS, or rather reflects an indirect relationship. For example, perhaps chemotherapy-induced immunosuppression acts as a risk factor for sepsis which, in turn, acts as a risk factor for ARDS.

Furthermore, the mortality associated with different clusters was also consistent with clinician experience. Clinicians recognize that patients with advanced malignancy may develop severe infections and ARDS as a final common pathway in their advanced disease. In this context, ARDS may represent a manifestation of their advanced underlying disease, and therapies directed at ARDS and its precipitant may not significantly impact mortality. Indeed, the cluster of malignancy-associated ARDS had a very high mortality rate of 90%.

In contrast, patients who develop ARDS as a result of trauma typically were well enough to be engaged in the activity leading to trauma, and thus ARDS may truly be the primary disease as opposed to a symptom of advanced underlying disease. The likelihood of mortality in this group may be more significantly influenced by the treatments targeted at ARDS. In our analysis, the cluster of trauma associated-ARDS had a significantly lower mortality rate of 30% compared to other clusters.

**Predictive Analytics**

Clustering and manual evaluation by clinicians of Patient Context Vectors proved to be a useful tool in summarizing the overall patient condition, risk factors and history. Real-time CDS systems, however, might not have access to the full set of the patient ICD codes as they might be entered in the EMR system at a later stage. It has been observed that clinical notes, specifically nursing and physician notes, typically contain all of the information available from ICD codes. Furthermore, while past medical history and presenting conditions might not always be ICD-coded, they are typically available in the form of free-text notes. In previous work\(^\text{30}\), we have successfully utilized free-text for predicting Patient Context Vectors, in the case of missing ICD codes. Additionally, we were able to successfully combine information available in ICD codes and in nursing notes and produce an average Patient Context Vector.

A word-level Convolutional Neural Network (CNN) was trained to predict from free-text notes the Patient Context Vector (averaged ICD code embedding). Structured data, in the form of vital signs and lab results, was then combined with the predicted Patient Context Vector and utilized in a machine learning model trained to predict ARDS patients.
Table 1: The top 15 most representative ICD codes for various clusters, based on cosine similarity to the cluster centroid.

Figure 3 summarizes the system workflow during prediction time.

The ARDS prediction model was trained utilizing structured data from the first 24 hours of admission, as described above in the Dataset Section. In addition, the first half of the available patient nursing notes and the first half of entered ICD codes were used to produce Patient Context Vectors of size 50. The patient ICD codes were used to look
Figure 3: Real-time ARDS prediction workflow. Nursing notes available at prediction time are used to predict Patient Context Vectors. ICD codes available at prediction time are also converted to Patient Context Vectors by averaging ICD code embeddings. Patient Context Vectors are used together with structured EMR data to predict the patient ARDS status.

<table>
<thead>
<tr>
<th>ARDS Prediction Features</th>
<th>AUC</th>
<th>P</th>
<th>R</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>79.74</td>
<td>29.77</td>
<td>67.80</td>
<td>37.17</td>
</tr>
<tr>
<td>Baseline + first half of notes/ICD</td>
<td><strong>90.16</strong></td>
<td><strong>49.46</strong></td>
<td><strong>78.85</strong></td>
<td><strong>48.63</strong></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>ARDS Mortality Prediction Features</th>
<th>AUC</th>
<th>P</th>
<th>R</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>78.26</td>
<td>69.19</td>
<td>92.60</td>
<td>79.20</td>
</tr>
<tr>
<td>Baseline + first half of notes/ICD</td>
<td><strong>82.11</strong></td>
<td><strong>73.66</strong></td>
<td><strong>89.99</strong></td>
<td><strong>81.01</strong></td>
</tr>
</tbody>
</table>

Table 2: 10-fold cross-validation GBM results of predicting ARDS patients and predicting mortality among ARDS patients. P=Precision, R=Recall, F1= F1-score for the positive class. The Baseline set of features consists of vital signs, lab results, Glasgow Coma Scale score, gender and age, in the form of structured data. “Baseline + first half of notes/ICD” includes also the average of the first half of entered visit ICD codes embeddings, and Patient Context Vectors predicted from the first half of the visit nursing notes.

A Gradient Boosting Machine (GBM) model\textsuperscript{31,32} was used to predict ARDS patients from the total population of adult patients. A GBM model was also used to predict the mortality among all ARDS patients. Table 2 shows the result from the experiments. All results were produced via 10-fold cross validation.

In both prediction models, the inclusion of Patient Context Vectors significantly increased the overall model performance. Intuitively, the patient medical history and overall clinical condition are important predictive factors. Results demonstrate that Patient Context Vectors can be successfully utilized to represent additional knowledge of a patient’s condition. The importance of the Patient Context Vector is also demonstrated by the scaled GBM variable importance shown in Figure 4. Together with critical vital signs measurements, various Patient Context Vector dimensions (shown with prefix embedding) play an important role in predicting the patient’s ARDS outcome. A known limitation of
low-dimensionality representations is the lack of interpretably of individual dimensions (e.g. embedding 39 lacks the interpretability of systolic blood pressure or temperature). Future work will focus on interpretability-imparted patient context vector embeddings, amenable to clinical interpretation.

In terms of practical application, the proposed system can be used in addition to existing high recall/low precision hospital alert systems, and used to prioritize alerts, mitigating the effects of alert fatigue. Furthermore, the imperfect and noisy nature of the automatically created dataset is likely resulting in over-pessimistic evaluation. It has been shown that ML classification algorithms are able to achieve high performance at relatively high levels of noise and that ML models generated from noisy datasets perform significantly better when evaluated on clean test sets (10 to 30% classification accuracy improvement at high training set noise levels)\textsuperscript{33–35}. Future work will focus on creating a clean, clinician-reviewed ARDS test dataset for more precise evaluation of the proposed approach.

**Related Work**

Numerous prediction scores have been developed to assess ARDS prognosis. Gajic et al.\textsuperscript{5} developed a Lung Injury Prediction Score (LIPS) formula including predisposing conditions, such as sepsis, shock, pneumonia, alcohol abuse, chemotherapy, FIO\textsubscript{2} and respiratory rate measures, found useful in predicting ARDS and mortality in surgical critical care patients\textsuperscript{36}. Interestingly, over 70% of score points associated with LIPS score calculation are based on patient context data rather than vials, labs, symptoms.

Other tools, such as Villar et al.\textsuperscript{6} base their ARDS prediction score on Age, PaO\textsubscript{2}/FIO\textsubscript{2}, and Plateau pressure score. Levitt et al.\textsuperscript{7} developed an Early Acute Lung Injury (EALI) score including risk factors, respiratory rate, and oxygen requirement. Xie et al.\textsuperscript{8} developed a modified ARDS prediction score (MAPS) based on a hand-crafted set of risk factors, risk modifiers, vital signs, etc.

A number of studies focus on predicting mortality among ARDS patients. For example, similar to our cluster findings, Hyers\textsuperscript{37} reports that patients who develop bacterial sepsis and multiple organ dysfunction are at high risk of dying and patients who develop ARDS from trauma or other noninfectious causes have a better prognosis. Navarrete-Navarro et al.\textsuperscript{38} report that mortality among ARDS patients correlates with the PaO\textsubscript{2}/FIO\textsubscript{2} ratio on the 3rd day of ARDS, the APACHE III score, and the development of multiple system organ failure. Timmons et al.\textsuperscript{39} studied children with ARDS and found significant differences between survivors and nonsurvivors who develop ARDS from trauma or other noninfectious causes. Timmons et al.\textsuperscript{39} studied children with ARDS and found significant differences between survivors and nonsurvivors who develop ARDS from trauma or other noninfectious causes. Timmons et al.\textsuperscript{39} studied children with ARDS and found significant differences between survivors and nonsurvivors who develop ARDS from trauma or other noninfectious causes.

Unlike the current work, the described previous studies utilized only structured patient data, and ICD codes/risk factors, when used, consisted of manually crafted lists.

In a broader context, a large volume of literature on combining structured and free-text EMR data apply Medical Concept detection on the free-text notes for manually curated list of risk factors and other disease-relevant medical
concepts. Ford et al.\textsuperscript{42} present a review of various approaches to Medical Concept detection from free-text notes for the purpose of detecting cases of a clinical condition, often in conjunction with structured data.

More recently, deep learning has been used to utilize free-text and structured EMR data. Shickel et al.\textsuperscript{43} present a survey of various deep learning techniques. Miotto et al.\textsuperscript{44} build a Deep Patient representation in an unsupervised manner via denoising autoencoders, however, similar to previous approaches they first pre-process the free-text notes by extracting medical concepts with an off-the-shelf tool. Various studies\textsuperscript{25–27, 45} use deep learning techniques to generate low-dimensional representations of diagnosis codes and patients utilizing structured data (diagnosis codes, medications, and procedures). Unlike previous work, we combine free-text and structured EMR data for obtaining low-dimensional patient representations, without the use medical concept detection.

Conclusion
This work demonstrates the utility of deep learning techniques to summarize a patient’s medical history, risk factors, comorbidities, and current signs and symptoms in the form of Patient Context Vectors. Automatically generated ARDS patients clusters agree with manually curated clinician knowledge and provide additional insight into the complexities and risk factors associated with ARDS. More importantly, Patient Context Vectors, derived from available ICD codes and nursing notes, can be easily combined with structured EMR data to build real-time ARDS CDS tools, with potential to improve patient outcomes and reduce mortality among ARDS patients.

Acknowledgements
Research reported in this publication was supported by a NIH SBIR award to CTA by NIH National Heart, Lung, and Blood Institute, of the National Institutes of Health under award number 1R43HL135909-01A1.

References


