A Computable Phenotype for Acute Respiratory Distress Syndrome Using Natural Language Processing and Machine Learning

Majid Afshar, MD, MSCR[1,2], Cara Joyce, PhD[2], Anthony Oakey, BS[3], Perry Formanek MD[4], Philip Yang, MD[4], Matthew M. Churpek, MD, MPH, PhD[5], Richard S. Cooper, MD[2], Susan Zelisko, MS[6], Ron Price, BS[6], Dmitriy Dligach, PhD[2,3]

[1]Division of Pulmonary and Critical Care Medicine, Loyola University Medical Center, Maywood, IL; [2]Department of Public Health Sciences, Stritch School of Medicine, Loyola University Chicago, Maywood, IL; [3]Department of Computer Science, Loyola University Chicago, Chicago, IL; [4]Department of Medicine, Loyola University Medical Center, Maywood, IL; [5]Division of Pulmonary and Critical Care Medicine, University of Chicago, Chicago, IL; [6]Informatics and Systems Development, Health Sciences Division, Loyola University Chicago, Maywood, IL

Abstract

Acute Respiratory Distress Syndrome (ARDS) is a syndrome of respiratory failure that may be identified using text from radiology reports. The objective of this study was to determine whether natural language processing (NLP) with machine learning performs better than a traditional keyword model for ARDS identification. Linguistic pre-processing of reports was performed and text features were inputs to machine learning classifiers tuned using 10-fold cross-validation on 80% of the sample size and tested in the remaining 20%. A cohort of 533 patients was evaluated, with a data corpus of 9,255 radiology reports. The traditional model had an accuracy of 67.3% (95% CI: 58.3-76.3) with a positive predictive value (PPV) of 41.7% (95% CI: 27.7-55.6). The best NLP model had an accuracy of 83.0% (95% CI: 75.9-90.2) with a PPV of 71.4% (95% CI: 52.1-90.8). A computable phenotype for ARDS with NLP may identify more cases than the traditional model.

Introduction

Acute respiratory distress syndrome (ARDS) is a common manifestation of pulmonary organ failure and a syndrome with profound hypoxemia with a period prevalence of 10% in all intensive care unit (ICU) admissions, and an associated mortality as high as 46% for those with severe ARDS.[1] Recognition by the clinician occurs in only 34% of cases once ARDS criteria are met.[1] One reason why ARDS is underrecognized by physicians is because it is a complex definition that incorporates laboratory data, respiratory data, radiology data, and disease characteristics within a time-sensitive framework.[2]

Automated methods for ARDS have served as one solution for syndrome detection.[3-5] A prior publication demonstrated some success in a hand-crafted, rule-based approach using keywords in dictated chest radiograph reports. This *traditional model* identified keywords that were synonyms for 'bilateral' and 'infiltrate' in the chest radiograph reports to meet the imaging parameters of the ARDS definition.[2] In follow-up studies, the traditional model has not performed well with high false positive rates.[6,7] Differences in sub-phenotypes of ARDS such as trauma and medical patients pose one barrier in validity.[8] In addition, keywords in a chest radiograph derived for an automated approach may vary between centers and cohorts making scalability a challenge.

Clinical notes and reports entered by healthcare providers are the most abundant type of data in the electronic health record (EHR) and serve as valuable sources of patient information[9,10]. Natural language processing (NLP) is a field of computer science and artificial intelligence that is concerned with developing methods for computational analysis of human (natural) language, including clinical notes.[11] NLP is frequently utilized in combination with machine learning, which involves extracting text-based features and training a machine learning classifier. Researchers have evaluated the performance of NLP for ARDS identification.[6, 7] However, these studies referenced against the older American European Consensus Conference definition instead of the current Berlin Definition[2], only used chest

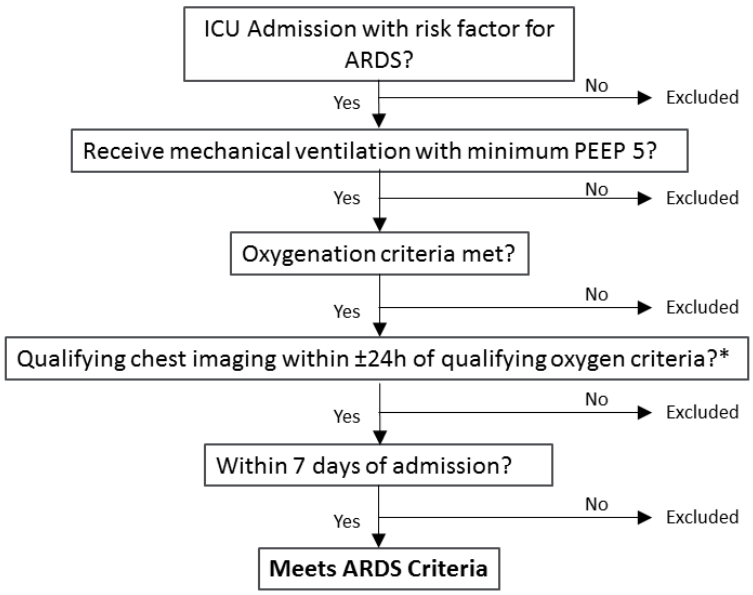
radiograph reports and did not include computed tomography (CT) reports, and did not examine standardized terminology as text-based features. The benefits of NLP and machine learning models have not been evaluated for identification of ARDS in accordance with the Berlin Definition, in multiple sub-phenotypes, or with different text-based features.

For the first time, we aim to evaluate the test characteristics of NLP and machine learning models in a mixed cohort of medical, trauma, and burn patients following the Berlin Definition utilizing multiple radiology reports (chest radiographs and CTs) and incorporating both raw text and standardized terminology. We hypothesize that NLP and machine learning can better identify ARDS cases than the traditional model.

Methods

Patient Selection and Environment:
A mixed cohort of 533 patients admitted to the medical, burn, and trauma ICUs at a tertiary academic center between January 1, 2011 and December 31, 2016 were included in the analysis. Inclusion criteria were the following: (1) ICU admission; (2) invasive mechanical ventilation with a minimum positive end expiratory pressure (PEEP) of 5 cm H_2O; (3) at least one partial pressure of arterial blood-to-fraction of inspired oxygen (PaO2/FiO2) <300 mmHg. A simple random sample was selected to represent a heterogenous group of risk factors for development of ARDS. The medical ICU group was part of a parent study examining risk factors for ARDS in critically ill patients with cirrhosis. Patients admitted to the burn and trauma ICUs for non-burn/trauma injuries and subsequent encounters for prior injuries were excluded. Burn-injury was verified after patient discharge by review of operative notes and attending documentation. All patients receiving invasive mechanical ventilation with suspected inhalation injury had a bronchoscopy to confirm inhalation injury based on the Abbreviated Injury Score.[12] Injury severity scores for trauma patients were verified by the trauma registry coders.

Figure 1.*Traditional model for ARDS identification*



*

Keywords: all combinations between location and morphology examined AND/OR selected words	Location: bilateral, biapical, bibasilar, widespread, diffuse, perihilar, parahilar, multifocal, extensive Morphology: infiltrates, opacities, air and space, aspiration, consolidation, pneumonia Selected keywords: congestive heart failure, left and right sided infiltrates, ARDS, pulmonary edema

Qualifying chest imaging includes chest radiograph and computed tomography. Oxygenation criteria is first qualifying partial pressure of arterial blood-to-fraction of inspired oxygen (PaO2/FiO2) <300 mmHg
PEEP = positive end-expiratory pressure in cm H_20

Reference Standard for ARDS

Two blinded physician reviewers (PY, PF) independently annotated cases and non-cases of ARDS in agreement with the Berlin Definition. The reviewers were internal medicine residents and received training to implement an ARDS screen in the ICU as part of a quality improvement project and reviewed cases prospectively with attending ICU physicians over a three-month period. The inter-rater agreement between the two reviewers was substantial, with a κ-coefficient at 0.71 (95% confidence interval [CI]: 0.60-0.82). A third physician with board certifications in pulmonary and critical care as well as research experience in ARDS (MA) provided adjudication for discordant cases.[13-15] In all cases of ARDS, both reviewers had to agree that respiratory failure was not fully explained by cardiac failure or hydrostatic pulmonary edema in accordance with the requirements stated in the Berlin Definition.[2]

Traditional model (modified to Berlin Definition)

A previously published hand-crafted, keyword method was adapted to the Berlin definition in a rule-based algorithm with keywords determined *a priori* **(Figure 1**).[4,5] The algorithm was updated from the original studies to include CT reports, minimum PEEP of 5 cm H_2O, and within seven days of hospital presentation. The qualifying keywords for bilateral infiltrates matched the prior publications except the word "consolidation" was added as a qualifier because it is a common term in CT reports.

Feature extraction

Linguistic processing of clinical notes was performed in clinical Text Analysis and Knowledge Extraction System (cTAKES; http://ctakes.apache.org).[16] Two methods of feature extraction were performed. In the first method, the spans of Unified Medical Language System (UMLS) named entity mentions (diseases/disorders, signs/symptoms, anatomical sites, procedures) were identified. Each Named Entity mention was mapped to a UMLS concept unique identifier (CUI) to standardize language variation between radiologists. For example, a mention of 'ARDS' in the text of the note is mapped to CUI C0035222; mentions of 'wet lung', 'adult respiratory distress syndrome', etc. are mapped to the same unique identifier. Each named entity mention was subsequently analyzed to determine its negation (e.g., 'no edema vs 'edema') status using the cTAKES negation module. For the second method, word n-grams (sequence of adjacent words of length n) were evaluated as a separate feature type. Both feature types are commonly used for text classification in clinical and general domains.[17] A term-frequency, inverse document-frequency (tf-idf) transformation was used to weigh the text features into normalized values for the machine learning classifiers.

Analysis Plan with Supervised Machine Learning

Descriptive statistics between ARDS and non-ARDS cases were examined. Continuous variables were evaluated as medians and interquartile ranges and analyzed with Wilcoxon rank sum test. Categorical variables were analyzed using chi-square tests.

The primary analysis was the binary classification of ARDS using radiology reports within 24 hours of the first qualifying PaO2/FiO2 ratio. This matched the traditional model that was previously published.[4,5] A model using all radiology reports was examined as well. The sample size was divided into 80% training and 20% testing for all machine learning classifiers. Text features (CUIs vs. n-grams) were inputs to separate machine learning classifiers, and classifier hyperparameters were tuned to the highest area under the receiver operating characteristic curve (AUC ROC) using 10-fold cross-validation on the training dataset. The best model from the training dataset was tested in the remaining 20% of the sample. We evaluated multiple machine classifiers as implemented in Scikit-Learn Version 0.19.0 (http://scikit-learn.org/) including decision trees, k-nearest neighbors, naíve bayes, logistic regression, and support vector machine (SVM). In prior NLP work, SVM with a linear kernel has traditionally proven to be the optimal algorithm for document classification.[18,19]

Discrimination of the prediction models was evaluated using the AUC ROC. Goodness-of-fit was formally assessed by the Hosmer-Lemeshow test and verified visually with a calibration plot. Test characteristics (accuracy, sensitivity, specificity, negative predictive value (NPV), and positive predictive value (PPV)) were provided to compare the traditional rule based model to the machine learning models. The nonparameteric approach by DeLong et al.[20] was used to compare the area under the ROC curves. Analysis was performed using Python Version 2.7.14 and SAS Version 9.4 (SAS Institute, Cary, NC). The Institutional Review Board of Loyola University Chicago approved this study. This research was supported in part by the National Institute of Health (K23AA024503).

Results

Patient and data characteristics

The data corpus was comprised of 162,440 radiology reports and clinical notes from 533 patients admitted to three ICUs. The count decreased to 9,255 for radiology reports only (any chest radiograph or CT that involved the chest) and 1,704 for the radiology reports within 24 hours of the first qualifying PaO2/FiO2 ratio. In the mixed cohort of patients admitted to the medical, trauma, and burn ICUs, the case-rate for ARDS was 25.9% (N=138). The median injury severity score in trauma patients was severe at 15 (IQR 9-22). Burn patients had a median % total body surface area burn injury at 19.7% (IQR 9.2%-40%), and 35.5% (N=55) had inhalation injury. In the patients admitted to the medical ICU, 57.9% (N=114) had sepsis. The lowest proportion of patients with ARDS was in those with trauma, and the greatest proportion in burn-injured patients (**Table 1**). Patients with ARDS had worse PaO2/FiO2 ratios, and a greater proportion died in-hospital than their non-ARDS counterparts (**Table 1)**. During chart annotations of the medical ICU subgroup, only 31% (N=14) of the patients with ARDS had any mention of ARDS in the EHR clinical notes or radiology reports.

Table 1. *Patient characteristics and outcomes*

	Total (N=533)	**No ARDS (n=395)**	**ARDS (n=138)**	**p-value**
Age, median (IQR)	55 (41-65)	56 (41-66)	53 (41-62)	0.21
Sex, male, n (%)	367 (68.9)	274 (69.4)	93 (67.4)	0.67
Race, white, (%)	349 (65.5)	264 (66.8)	85 (61.6)	0.04
Charlson Comorbidity Index, median (IQR)	6 (3-7)	.36 (3-7)	5 (2-6)	0.36
Intensive Care Unit, n (%) Medical Trauma Burn	 197 (37.0) 181 (34.0) 155 (29.1)	 155 (39.2) 148 (37.5) 92 (23.3)	 42 (30.4) 33 (23.9) 63 (45.7)	<0.001
Pa02/Fi02 ratio, median (IQR)	211.9 (146.5-262.0)	225.0 (163.3-265.0)	169.0 (121.4-242.0)	<0.001
Hospital length of stay (days), median (IQR)	16.2 (7.6-26.9)	15.4 (7.5-23.5)	22.0 (7.6-39.6)	<0.001
Disposition, n (%) In-Hospital Death Acute/Subacute Care Chronic Care Home Other	 175 (32.8) 129 (24.2) 53 (9.9) 148 (27.8) 28 (5.3)	 102 (25.8) 101 (25.6) 38 (9.6) 128 (32.4) 26 (6.36)	 73 (52.9) 28 (20.3) 15 (10.9) 20 (14.5) 2 (1.4)	<0.001

Acute/subacute care= skilled nursing facility, inpatient rehab; Chronic Care= long term care hospital, nursing facility,
Other = hospice, against medical advice, law enforcement, psychiatric hospital; Charlson Comorbidity Index is a comorbidity score used to predict 10-year mortality.

Discrimination and Calibration of NLP and Machine Learning Models:

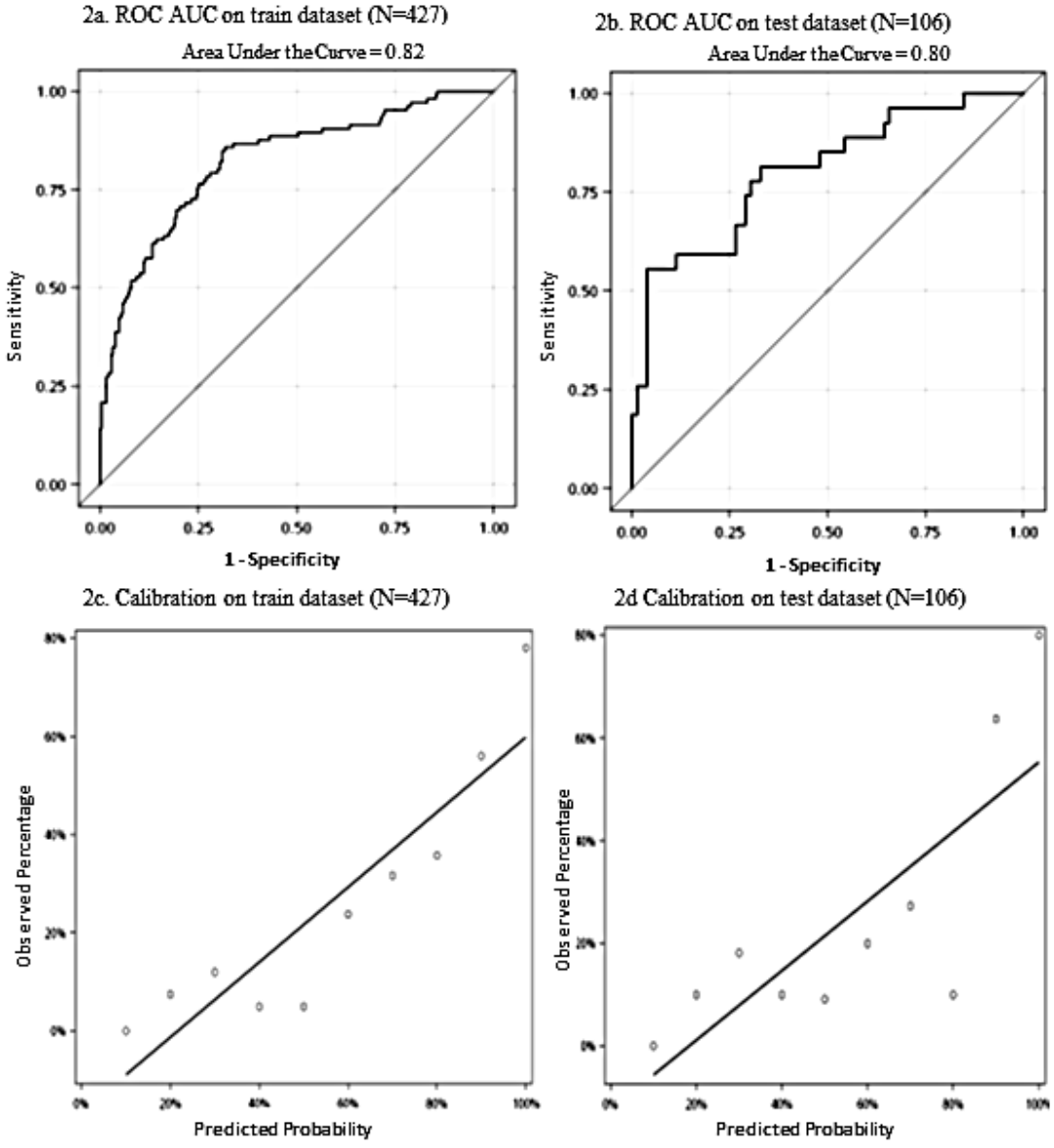
In the same approach as the traditional method using radiology reports within 24 hours of qualifying PaO2/FiO2 ratio, the NLP model that produced the best AUC ROC was a linear SVM classifier with unigram features with an AUC in the test dataset of 0.73 (95% CI: 0.61 – 0.85) (**Table 2)**. The addition of all radiology reports was examined and a linear SVM classifier with unigram features was derived with an increase in AUC ROC to 0.81 (95% CI: 0.72-0.91) in the test dataset**.** In the train dataset, bi-gram features did not improve the AUC ROCso they were not evaluated further (data not shown). Using CUIs instead of unigram features in the model with all radiology reports did not produce a significant change in discrimination by AUC ROC (p=0.24).

Table 2. *Area Under the Receiver Operating Characteristic Curve for Different NLP models*

NLP and Machine Learning Model	Train Dataset (95% Confidence Interval)	Test Dataset (95% Confidence Interval)
	N=427	N=106
24hr radiology reports and unigrams	0.82 (0.77-0.86)	0.73 (0.61-0.85)
All radiology reports and unigrams	0.84 (0.80-0.89)	0.81 (0.72-0.91)
All radiology reports and CUIs	0.82 (0.78-0.87)	0.80 (0.70-0.90)

Discrimination of the NLP model using all radiology reports and CUI features is shown with the AUC ROC curve in **Figure 2a**, and **Figure 2b** is the corresponding calibration plot. The Hosmer-Lemeshow Goodness-of-Fit test showed the NLP model with all radiology reports and CUI features fit the test dataset well (p=0.25).

Figure 2a-b. *Discrimination with Area Under the Receiver Operative Characteristic Curve and Calibration plots of NLP model with all radiology reports and Concept Unique Identifier features*



Comparison between traditional model and NLP model

The traditional model had an overall accuracy of 67.3% (95% CI: 58.3-76.3) with a PPV of 41.7% (95% CI: 27.7-55.6). The remaining test characteristics are shown in **Table 3** with comparisons to all three NLP model variations. Overall, an increase in accuracy and PPV occurs by expanding to a data corpus of all radiology reports and using CUIs instead of unigram features in a linear SVM classifier.

Table 3. *Test Characteristics of selected algorithms in the test dataset*

Model	Accuracy (95% CI)	Sensitivity (95% CI)	Specificity (95% CI)	PPV (95% CI)	NPV (95% CI)
Traditional: 24hr radiology reports and keywords	67.3% (58.3-76.3)	76.9% (60.7-93.1)	64.1% (53.5-74.8)	41.7% (27.7-55.6)	89.3% (81.2-97.4)
NLP + machine learning: 24hr radiology reports and unigrams	76.9% (68.8-85.0)	42.3% (23.3-61.3)	88.5% (81.4-95.6)	55.0% (33.2-76.8)	82.1% (74.0-90.3)
NLP + machine learning: All radiology reports and unigrams	80.2% (72.6-87.8)	55.6% (36.8-74.3)	88.6% (81.6-95.6)	62.5% (43.1-81.9)	85.4% (77.7-93.0)
NLP + machine learning: All radiology reports and CUIs	83.0% (75.9-90.2)	55.6% (36.8-74.3)	92.4% (86.6-98.3)	71.4% (52.1-90.8)	85.9% (78.5-93.3)

Each classifier was examined in the test set of N=106 patients. All NLP algorithms were performed in a linear support vector machine classifier. PPV = positive predictive value; NPV = negative predictive value; CUI = concept unique identifier.

Characteristics of n-grams and CUIs as text features

In examining all radiology reports, there were 1,774 unique CUIs and the top positive weights for ARDS cases included "ARDS", "aspiration", and "Both lungs". The top negative weights for non-ARDS cases included "atelectasis", "pneumothorax", and "chronic obstructive pulmonary disease". The top 25 positive and negative weights from the linear SVM model for CUIs are listed in **Table 4.**

Table 4. *Top features from NLP model using all radiology reports (negative weights favor the non-ARDS cases and positive weights favor ARDS cases)*

Positive Concept Unique Identifier (CUI) features (weights from linear support vector machine classifier)	Respiratory Distress Syndrome, Adult (0.22); Malaise (0.21); Fluid overload (0.19); Aspiration-action (0.17); Fundus (0.16); Both lungs (0.15); Communicable diseases (0.15); Pulmonary edema (0.13); Lung field (0.13); Infiltration (0.12); learn transplantation (0.12); Air (0.11); Cardia of stomach (0.11); Inhalation injury (0.11); Rest (0.10); On ventilator (0.10); Structure of subclavian vein (0.10); Disease progression (0.10); Reflecting (0.10); Severe pre-eclempsia (0.10); Veins (0.09); Surface region of upper chest (0.09); Liver cirrhosis (0.09); Hyperemia (0.08); Closure by staple (0.08)
Negative Concept Unique Identifier (CUI) features (weights from linear support vector machine classifier)	Dose pull (0.09); Edema (0.08); Structure of carina (0.08); Transplantation of liver (0.08); Arteries (0.08); plain chest x-ray (0.08); pneumothorax (0.08); tracheal extubation (0.08); transjugular intrahepatic portosystemic shunt procedure (0.07); intestinal obstruction (0.07); nasogastric feeding (0.06); trachea (0.06); neck (0.06); chronic obstructive airway disease (0.06); upper abdomen structure (0.06); gas dosage form (0.06); fever (0.06); probably present (0.05); head (0.05); atelectasis (0.05); widened mediastinum (0.05); skin appearance normal (0.05); gait normal (0.05); comminuted fracture type (0.05); limb structure (0.05)

Discussion

ARDS is a syndrome with many predisposing conditions and a definition that incorporates both structured and unstructured elements rendering a complex phenotype that is difficult to automate. This highlights the difficulty of the task; in absence of direct evidence of ARDS, a machine learning model can incorporate other relevant features to make the identification of ARDS. We demonstrated a scalable and useful approach for a computable phenotype of ARDS using NLP and supervised machine learning. The NLP model using standardized terminology with CUIs

from all radiology reports increased the PPV from 41.7% to 71.4% and improved the accuracy from 67.3% to 83.0% over the traditional model.

We examined a cohort of patients at high-risk for ARDS with a mixture of predisposing conditions. Inhalation injury was among the predisposing conditions and has previously been shown in population-based studies to carry the highest incidence for ARDS.[21] Severe trauma followed as one of the next highest risk categories.[21] We identified severely injured burn and trauma patients in an attempt to provide a more balanced dataset of ARDS and non-ARDS cases to improve model performance. Furthermore, the inclusion of decompensated cirrhosis in the medical ICU cohort introduced additional respiratory complications such as hydrostatic pulmonary edema and pleural effusions that are common to this group.[22] Therefore, we identified cohorts that not only provided multiple etiologies for respiratory failure but also represented a high-risk heterogeneous group of direct and indirect lung injuries to address the complexities of the syndrome.

During annotation, we observed that clinical notes rarely mentioned ARDS. Others have also demonstrated poor reliability from claims data for identifying ARDS.[23] To address this problem, the traditional model was derived by a group of investigators using a cohort of patients with severe trauma from a single site.[4] This original study had a PPV of 74% using a cohort of 199 patients with 53 (26%) cases of ARDS. A separate validation study screened 1,270 patients and identified 84 (6.6%) cases of ARDS with a PPV of 74%.[5] However, follow-up studies from other sites have shown worse performance using the same traditional model with a lower PPV around 40%.[6, 7] Our cohort of 533 patients with a quarter having ARDS had similar results to the follow-up studies with a PPV of 41.7%.

In the two follow-up studies that examined NLP versus the traditional model for ARDS phenotyping,[6, 7] the authors focused on n-grams using solely raw text available in chest radiograph reports. They had accuracies between 73% and 91% with PPVs better than the traditional model. We noted similar results but validation of unigram features in radiology reports within 24 hours of the qualifying oxygenation report demonstrated a substantial drop in AUC ROC in the test dataset, suggesting the model may be overfit. However, incorporating all radiology reports provided AUC ROCs above 80% in both train and test datasets. We showed that not only was discrimination good but model fit was adequate when visualized by the calibration plot. Although CUIs did not demonstrate better discrimination over unigram features, their use as a standardized terminology have better ability to scale across hospital sites. In future work, we plan to conduct external validation at another site.

Many of the top weighted CUI features from the linear SVM model were similar to the keywords from the traditional model and are consistent with the clinical knowledge. CUIs such as ARDS, both lungs, infiltration, and pulmonary edema were among the similarities. Edema as a top feature for both the positive CUIs (fluid overload and pulmonary edema) as well as negative CUIs (edema) likely reflected the poor inter-observer reliability between dictated radiology reports for cases of ARDS.[24] Uncertainty in labelling of cases versus non-cases of ARDS remains a problem with poor inter-observer reliability during annotation.[25] This was recently addressed by a study from Reamaroon et al. that accounted for label uncertainty and subsequently improved their AUC ROC from 75% to 85% for discrimination of ARDS.[26] In addition, atelectasis and heart failure CUIs were important negative features that support reasons for hypoxemia in non-ARDS cases. Some of the top negative CUIs were representative of other diseases associated with hypoxemia such as pneumothorax and chronic obstructive pulmonary disease. Certain features in our model was also unique to the sub-phenotypes of ARDS. For instance, some of the top features for ARDS included concepts unique to trauma and burn injury. The features unique to sub-phenotypes of ARDS suggest that computable phenotypes may need to be tailored more for specific cohorts rather than an all-inclusive model. Our sample size did not allow for adequate analyses by sub-phenotype to investigate this further.

The use of all radiology reports may not be appropriate for an early detection screening tool but is useful for a surveillance tool in epidemiologic studies. The inclusion of all radiology reports over the 24-hour reports substantially increased the data corpus and provided the best PPV and overall accuracy. Although the NLP model with all radiology reports may serve as a useful research tool to extract ARDS cases from the EHR, a more time-sensitive model with 24-hour reports for early detection has better clinical application. In future work, we plan to investigate why the latter model underperformed in our experiments to improve its test characteristics. In addition, we plan to train word embeddings using *word2vec* to implement a convolutional neural network model and examine subsumptions in our CUI-based approach to improve our results further.

Limitations of this study include its retrospective and single site design. The heterogeneity of the dataset with different sub-phenotypes also likely contributed to the lower sensitivities compared to other studies. In addition, the use of all radiology reports may have introduced additional error if a qualifying oxygenation ratio was not temporally linked to the features in the notes to fulfill the Berlin Definition. However, we did not find that test characteristics worsened when comparing radiology reports within 24 hours versus those not within 24 hours of the first qualifying oxygen ratio. This may also have been attributable to additional qualifying oxygen ratios that were not extracted from the structured data elements.

Conclusion

This study presents useful approaches from NLP and machine learning to build a computable phenotype for ARDS. To our knowledge, this is the first study using advanced feature engineering with linkage to UMLS for standardized terminology with CUIs, incorporating all radiology note types, and updating to the Berlin Definition. With additional improvements in this approach and external validation, large-scale implementation across EHRs for identification and surveillance is feasible and may inform future research studies and practice. Moving forward we plan to experiment with neural network models and ensemble models combining classifiers trained on both structured and unstructured data.

References:

1. Bellani G, Laffey JG, Pham T, Fan E, Brochard L, Esteban A, Gattinoni L, van Haren F, Larsson A, McAuley DF, Ranieri M, Rubenfeld G, Thompson BT, Wrigge H, Slutsky AS, Pesenti A, Investigators LS and Group ET. Epidemiology, Patterns of Care, and Mortality for Patients With Acute Respiratory Distress Syndrome in intensive care units in 50 Countries. JAMA 2016;315:788-800.
2. Ranieri VM, Rubenfeld GD, Thompson BT, Ferguson ND, Caldwell E, Fan E, Camporota L and Slutsky AS. Acute respiratory distress syndrome: the Berlin Definition. JAMA 2012;307:2526-33.
3. Herasevich V, Yilmaz M, Khan H, Hubmayr RD and Gajic O. Validation of an electronic surveillance system for acute lung injury. Intensive Care Medicine 2009;35:1018-1023.
4. Koenig HC, Finkel BB, Khalsa SS, Lanken PN, Prasad M, Urbani R and Fuchs BD. Performance of an automated electronic acute lung injury screening system in intensive care unit patients. Crit Care Med 2011;39:98-104.
5. Azzam HC, Khalsa SS, Urbani R, Shah CV, Christie JD, Lanken PN and Fuchs BD. Validation Study of an Automated Electronic Acute Lung Injury Screening Tool. J Am Med Inform Assoc 2009;16:503-508.
6. Solti I, Cooke CR, Xia F and Wurfel MM. Automated classification of radiology reports for acute lung injury: Comparison of keyword and machine learning based natural language processing approaches. Proceedings (IEEE Int Conf Bioinformatics Biomed) 2009;2009:314-319.
7. Yetisgen-Yildiz M, Bejan CA and Wurfel MM. Identification of patients with acute lung injury from free-text chest x-ray reports. Proceedings of the 2013 Workshop on Biomedical Natural Language Processing (BioNLP 2013) 2013:10-17.
8. Calfee CS, Janz DR, Bernard GR, May AK, Kangelaris KN, Matthay MA, Ware LB and Network NNA. Distinct molecular phenotypes of direct vs indirect ARDS in single-center and multicenter studies. Chest 2015;147:1539-48.
9. Castro VM, Dligach D, Finan S, Yu S, Can A, Abd-El-Barr M, Gainer V, Shadick NA, Murphy S, Cai T, Savova G, Weiss ST and Du R. Large-scale identification of patients with cerebral aneurysms using natural language processing. Neurology 2017;88:164-168.
10. Joffe E, Pettigrew EJ, Herskovic JR, Bearden CF and Bernstam EV. Expert guided natural language processing using one-class classification. J Am Med Inform Assoc 2015;22:962-6.
11. Gonzalez-Hernandez G, Sarker A, O'Connor K and Savova G. Capturing the patient's perspective: a review of advances in natural language processing of health-related text. Yearb Med Inform. 2017;26:214-227.
12. Endorf FW and Gamelli RL. Inhalation injury, pulmonary perturbations, and fluid resuscitation. J Burn Care Res 2007;28:80-3.
13. Afshar M, Smith GS, Terrin ML, Barrett M, Lissauer ME, Mansoor S, Jeudy J and Netzer G. Blood alcohol content, injury severity, and adult respiratory distress syndrome. J Trauma Acute Care Surg 2014;76:1447-55.

14. Afshar M, Netzer G, Mosier MJ, Cooper RS, Adams W, Burnham EL, Kovacs EJ, Durazo-Arvizu R and Kliethermes S. The Contributing Risk of Tobacco Use for ARDS Development in burn-injured adults with inhalation injury. Respir Care 2017.
15. Afshar M and Netzer G. The international epidemiology of acute respiratory distress syndrome: how can we think locally and measure globally? Crit Care Med 2014;42:739-40.
16. Savova GK, Masanz JJ, Ogren PV, Zheng J, Sohn S, Kipper-Schuler KC and Chute CG. Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. J Am Med Inform Assoc 2010;17:507-13.
17. Shivade C, Raghavan P, Fosier-Lussier E, Embi PJH, Elhadad N, Johnson SB, Lai Am. A review of approaches to identifying patient phenotype cohorts using electronic health records. J Am Med Inform Assoc 2014;21:221-230.
18. Muzaffar AW, Azam F and Qamar U. A Relation Extraction Framework for Biomedical Text Using Hybrid Feature Set. Comput Math Methods Med. 2015;2015:910423.
19. T. Joachims. Text categorization with support vector machines: Learning with many relevant features. Technical Report 23, Universität Dortmund, LS VIII, 1997.
20. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics 1988;44:837.
21. Gajic O, Dabbagh O, Park PK, Adesanya A, Chang SY, Hou P, Anderson H, Hoth JJ, Mikkelsen ME, Gentile NT, Gong MN, Talmor D, Bajwa E, Watkins TR, Festic E, Yilmaz M, Iscimen R, Kaufman DA, Esper AM, Sadikot R, Douglas I and Sevransky J. Early identification of patients at Risk of Acute Lung Injury: Evaluation of Lung Injury Prediction Score in a Multicenter Cohort Study. Am J Resp Crit Care Med 2011;183:462-470.
22. Hemprich U, Papadakos PJ and Lachmann B. Respiratory failure and hypoxemia in the cirrhotic patient including hepatopulmonary syndrome. Curr Opin Anaesthesiol 2010;23:133-8.
23. Howard AE, Courtney-Shapiro C, Kelso LA, Goltz M and Morris PE. Comparison of 3 methods of detecting acute respiratory distress syndrome: clinical screening, chart review, and diagnostic coding. Am J Crit Care 2004;13:59-64.
24. Rubenfeld GD, Caldwell E, Granton J, Hudson LD, Matthay MA. Interobserver variability in applying a radiographic definition for ARDS. Chest 1999;116:1347-1353.
25. Sjoding MW, Hofer TP, Co I, Courey A, Cooke CR, Iwashyna TJ. Interobserver reliability of the Berlin ARDS defition and strategies to improve the reliability of ARDS diagnosis. Chest 2018;153:361-367.
26. 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 Informatics 2018;1-9.