Open-source machine learning pipeline automatically flags instances of acute respiratory distress syndrome from electronic health records

Félix L. Morales[1], Feihong Xu[2], Hyojun Ada Lee[1], Helio Tejedor Navarro[1,3], Meagan A. Bechel[4,5], Eryn L. Cameron[6], Jesse Kelso[6], Curtis H. Weiss[1,7], Luís A. Nunes Amaral[1,3,8,9,10]*

[1] Department of Engineering Science and Applied Mathematics, Northwestern University, Evanston, IL
[2] Interdepartmental Biological Sciences Program, Northwestern University, Evanston, IL
[3] Northwestern Institute on Complex Systems, Northwestern University, Evanston, IL
[4] Medical Scientist Training Program, Northwestern University Feinberg School of Medicine, Chicago, IL
[5] Department of Radiology, Emory University, Atlanta, GA
[6] Department of Medicine, Endeavor Health, Evanston, IL
[7] Division of Pulmonary and Critical Care Medicine, Endeavor Health, Evanston, IL
[8] Department of Physics and Astronomy, Northwestern University, Evanston, IL
[9] Department of Medicine, Northwestern University Feinberg School of Medicine, Chicago, IL
[10] NSF-Simons National Institute for Theoretical and Mathematical Biology, Chicago, IL

* Corresponding Author: Luís A. Nunes Amaral (amaral@northwestern.edu)

Abstract

Physicians could greatly benefit from automated diagnosis and prognosis tools to help address information overload and decision fatigue. Intensive care physicians stand to benefit greatly from such tools as they are at particularly high risk for those factors. Acute Respiratory Distress Syndrome (ARDS) is a life-threatening condition affecting >10% of critical care patients and has a mortality rate over 40%. However, recognition rates for ARDS have been shown to be low (30-70%) in clinical settings. In this work, we present a reproducible computational pipeline that automatically adjudicates ARDS on retrospective datasets of mechanically ventilated adult patients. This pipeline automates the steps outlined by the Berlin Definition through implementation of natural language processing tools and classification algorithms. First, we used labeled chest imaging reports from two different hospitals over three different time periods to train an XGBoost model to detect bilateral infiltrates, and a subset of attending physician notes from one hospital labeled for the most common ARDS risk factor (pneumonia) to train another XGBoost model to detect a pneumonia diagnosis. Both models achieve high performance when tested on out-of-bag samples—an area under the receiver operating characteristic curve (AUROC) of 0.88 for adjudicating chest imaging reports, and an AUROC of 0.86 for detecting pneumonia on attending physician notes. Next, we integrate the models and validate the entire pipeline on a fourth cohort from a third hospital (MIMIC-III) and find a sensitivity of 93.5% — an extraordinary improvement over the 22.6% ARDS recognition rate reported for these encounters — along with a false positive rate of 18.8%. We conclude that our reproducible, automated diagnostic pipeline exhibits promising ARDS retrospective adjudication performance, thus providing a valuable resource for physicians aiming to enhance ARDS diagnosis and treatment strategies. We surmise that real-time integration of the pipeline with EHR systems has the potential to aid clinical practice by facilitating the recognition of ARDS cases at scale.

Introduction

Physicians, including intensivists, process large amounts of dispersed information on many patients. This potential information overload poses serious risks to patient safety. Several studies[1–3] have estimated that 100,000-400,000 fatalities per year may be due to medical errors. While information overload is a challenge for humans, vast amounts of information become advantageous if used as an input for machine learning (ML) approaches. Recent advances in artificial intelligence, ML, and data science are enabling the development of protocols to extract knowledge from large datasets. However, some of those approaches lack interpretability and have been shown to be fragile (e.g., recent re-analysis of attempts to diagnose COVID-19 from chest X-ray images[4]).

In this study, we report the development and validation of a ML pipeline to help physicians adjudicate acute respiratory distress syndrome (ARDS). ARDS, a syndrome of severe acute hypoxemia resulting from inflammatory lung injury[5,6], is an ideal case for the development of a diagnostic aid tool. ARDS recognition requires physicians to synthesize information from multiple distinct data streams and determine whether it fits a standard definition of ARDS. The criteria of the clinically-based Berlin Definition of ARDS include quantitative data (p_aO_2:$F_IO_2 \leq 300$ mm Hg), unstructured data (bilateral opacities on chest imaging), and subjective data (assessing for the presence of ARDS risk factors and heart failure)[5]. Despite ARDS' high prevalence, morbidity, and mortality, prior research has shown that many patients with ARDS are not recognized by their treating physicians[6,7]. The poor recognition rate of ARDS — at times as low as 1 in 5[7] — is at least partially due to the difficulty in evaluating the Berlin Definition criteria, which requires the physician to access laboratory data, chest images or radiology notes, other physicians' notes, and echocardiographic data or notes, before they can apply the criteria to determine whether ARDS is present.

Under-recognition of ARDS plays an important role in under-utilization of evidence-based ARDS treatment (e.g., low tidal volume ventilation and prone positioning), even when physicians believe these interventions are warranted[8]. An automated approach to help identify ARDS diagnostic criteria has the potential to be a powerful aid to physician decision-making, leading to improved ARDS recognition and therefore improved ARDS management.

Previous studies have demonstrated some success in automating the recognition of individual ARDS diagnostic components using electronic health record (EHR) screening "sniffers"[9–11]. In addition, a ML algorithm to risk stratify patients for ARDS using structured clinical data derived from the EHR was shown to have good discriminative performance[12]. Regarding automating the entire ARDS diagnostic algorithm, two studies[13,14] have recently reported implementation of keyword search (i.e. rule-based approach) in the EHR with validation conducted for 100 intensive care unit (ICU) admissions from a single time period and from a single institution. A third study recently reported on a computable Berlin Definition which employed a previously developed neural network approach to adjudicate chest imaging reports restricted to patients with a single, known ARDS risk factor (COVID-19), with promising performance (93% sensitivity, 92% specificity)[15]. However, no study has successfully and reproducibly automated the entire sequence of steps required by the Berlin Definition of ARDS or tested the discriminative performance of the tool on a general population of critically ill patients who received invasive mechanical ventilation. Addressing this gap in a manner that overcomes considerable physician under-diagnosis of ARDS is the goal of this study.

Data and Methods

Cohort Development and Data Collection

We developed three patient cohorts for model training and evaluation: Hospital A (2013), Hospital A (2016), and Hospital B (2017-18), where A and B refer to two distinct and unaffiliated hospitals and the number in parentheses refers to the period over which data was collected. In addition, we obtained data from the Medical Information Mart for Intensive Care (MIMIC-III) database [16,17] MIMIC-III is a large, single-center database of critically ill patients at a tertiary care medical center. It includes all the components necessary to identify ARDS and therefore apply our pipeline, and the data is freely available. For naming consistency, we denoted the subset of MIMIC-III patients we consider as MIMIC (2001-12).

For all four cohorts, patients were included if they were at least 18 years old; were admitted to an adult ICU; and had acute hypoxemic respiratory failure requiring invasive mechanical ventilation (at least one recorded $P_aO_2/F_IO_2 \leq 300$ mm Hg while receiving positive-end expiratory pressure $\geq$ 5cm H_2O)[7]. The study was approved by the Institutional Review Boards of Northwestern University (STU00208049) and Endeavor Health (EH17-325). Table 1 summarizes the data collected for the four cohorts.

<u>Hospital A (2013)</u>

We previously characterized a cohort of 943 patients, which we denote here as Hospital A (2013), who met the above inclusion criteria at a single academic medical center between June and December 2013. We collected the following data: all P_aO_2/F_IO_2 ratios; the unstructured text of all radiologist reports for chest imaging (radiographs and CT scans), critical care attending physician notes, and echocardiogram reports; and B-type natriuretic peptide (BNP) values obtained from hospital admission to the earliest of extubation, death, or discharge. Data were reviewed by study personnel to determine whether each individual Berlin Definition criterion was present, and whether all criteria taken together were consistent with a diagnosis of ARDS [7].

We collected 5958 chest imaging reports from 800 Hospital A (2013) patients. Study personnel adjudicated 57% of these chest imaging reports as describing bilateral infiltrates consistent with the Berlin Definition[5]. We developed our machine learning (ML) approach to bilateral infiltrate adjudication using these Hospital A (2013) reports.

For 790 of the 800 Hospital A (2013) patients with a chest imaging report, we also had at least one attending physician note. We collected 12582 attending physician notes for these patients, of which 2034 notes from a subset of 400 patients were labeled by study personnel for the presence of ARDS risk factors (e.g., pneumonia, sepsis, aspiration, etc.)[5]. We used this labeled subset of 2034 notes to develop our ML and regular expression (regex) approach for finding ARDS risk factors and heart failure language in attending physician notes.

We collected 1006 echocardiogram (echo) reports from 681 Hospital A (2013) patients. Study personnel from a prior analysis[7] text-matched and adjudicated each echo report for the presence or absence of left ventricular ejection fraction < 40%, cardiopulmonary bypass at time of echo, left ventricular hypertrophy, left atrial dimension > 4cm or left atrial volume index > 28 mL/m^2, and Grade II or III diastolic dysfunction. Separately, 35 BNP values were included for 32 patients in Hospital A (2013). We used echo reports and BNP values to develop our objective heart failure rule-out approach.[7]

Since identification of Berlin Definition-consistent bilateral infiltrates is the most challenging task[11,18] in our computational pipeline, we obtained chest imaging reports from two additional cohorts of patients to strengthen our ML algorithm, Hospital A (2016) and Hospital B (2017-18).

Table 1. Cohort characteristics and data annotation. Hospital A, Hospital B, and MIMIC refer to three distinct and unaffiliated hospitals and the number in parentheses refers to the period over which data was collected.

Source	Report type	Number	% bilateral infiltrates	Rater IDs
Hospital A (2013)	Chest imaging reports	5958 reports (800 patients)	57%	1
Hospital A (2013)	Attending physician notes	12582 notes (790 patients)		
Hospital A (2013)	Adjudicated Attending physician notes	2034 notes (400 patients)		1, 5
Hospital A (2013)	Echocardiogram reports	1006 reports (681 patients)		
Hospital A (2016)	Chest imaging reports	6040 reports (749 patients)	44%	1
Hospital B (2017-18)	Chest imaging reports	625 reports (90 patients)	34%	1, 2, 3, 4
MIMIC (2001-12)	Chest Imaging reports	975 reports (100 patients)	22%	1,6
MIMIC (2001-12)	Attending physician notes	887 notes (100 patients)		1,6
MIMIC (2001-12)	Echocardiogram reports	89 reports (100 patients)		1,6

Hospital A (2016)

The Hospital A (2016) cohort comprises 749 patients admitted during 2016 at the same medical center as the Hospital A (2013) cohort and meeting the same inclusion criteria. We collected 6040 chest imaging reports for these patients. Study personnel adjudicated 44% of Hospital A (2016) reports as describing bilateral infiltrates. We also used these reports to develop our ML approach for bilateral infiltrate adjudication.

Hospital B (2017-18)

The Hospital B (2017-18) cohort comprises 90 patients admitted to a different, unaffiliated, medical center in 2017–2018 and meeting the same inclusion criteria as Hospital A (2013) and Hospital A (2016). We collected 625 chest imaging reports for these 90 patients. Study personnel adjudicated 34% of these chest imaging reports as describing bilateral infiltrates. Depending on the experiments (see below),

we used these reports from Hospital B (2017-18) to develop out ML algorithms for bilateral infiltrate adjudication, or to evaluate the robustness of XGBoost to changes in the training data.

MIMIC (2001-12)

We identified the set of patients in the MIMIC-III dataset who satisfied the inclusion criteria used to develop Hospital A (2013). This resulted in a set comprising 3712 encounters. We then used an earlier version of our automated pipeline to adjudicate the presence or absence of ARDS for all those encounters, and randomly selected a balanced cohort comprising 100 encounters, which we denote as the MIMIC (2001-12) cohort. Each of the encounters in MIMIC (2001-12) was adjudicated by one experienced critical care physician and one internal medicine physician for whether each individual Berlin Definition criterion was present, and whether all criteria taken together were consistent with a diagnosis of ARDS. We note that both reviewers were blinded to the individual encounter adjudications by the pipeline, although they did know that the ARDS split among the 100 encounters was more or less balanced. Prior to their review, the reviewers were blinded to the results of the automated pipeline. This cohort, with blinded physician adjudications, is publicly available at Northwestern's ARCH database).

The records of the 100 MIMIC (2001-12) patients included 975 chest imaging reports, 887 attending physician notes, and 89 echocardiogram (echo) reports. The critical care physician adjudicated 22.3% of these chest imaging reports as describing bilateral infiltrates consistent with the Berlin Definition[5]. The same individual also labeled 887 attending physician notes for the presence of ARDS risk factors[5] and heart failure language, and 89 echo reports for the presence or absence of left ventricular ejection fraction <40%, cardiopulmonary bypass at time of echo, left ventricular hypertrophy, left atrial dimension >4cm or left atrial volume index >28 mL/m^2, and Grade II or III diastolic dysfunction.[7] We used these adjudicated datasets as test to evaluate: (1) the performance of our implemented ML algorithms, and (2) the performance of the entire pipeline on a publicly available dataset.

Analysis

Adjudication of bilateral infiltrates from chest imaging reports

Overview

We used labeled chest imaging reports from Hospital A (2013), Hospital A (2016), and Hospital B (2017-2018) to develop our machine learning (ML) approach to adjudicating bilateral infiltrates in chest imaging reports. We will refer to this data as the "whole training set" — it comprises a total of 12,623 chest imaging reports of which 6272 (49.7%) reports were labeled positive for bilateral infiltrates. Since we employed cross-validation to estimate uncertainties in performance metrics, parts of the whole training set also served as a validation sets depending on the random splits done by the cross-validation routine.

In addition, we tested the robustness of the XGBoost classifier to changes in the training data by training two separate models: one trained on chest imaging reports from Hospital System A (2013), and another on chest imaging reports from Hospital System A (2016). We did not carry these models forward for use in the ARDS adjudication pipeline. Finally, we used all the 975 chest imaging reports from MIMIC (2001-12) as our test set.

Feature engineering

We preprocessed chest imaging reports by removing patient and clinician identification information, irrelevant sections (e.g. technique, indication, history, etc.), and 'non-informative' words. In selecting informative vs. non-informative words, we used the same inclusion and exclusion language criteria we previously developed to address limitations in Berlin Definition specificity[7]. We then tokenized the remaining sections — i.e., separated the text into sets of unigrams and bigrams — and prepared the data for use of a "bag of words" approach — i.e., vectorized these tokens according to their counts in the imaging reports. When training a ML model on a given corpus, we used the 200 most frequently appearing tokens across the imaging reports in the respective corpus as model features.

Training of models for adjudication of bilateral infiltrates

We compared the performance of four different classification algorithms — decision trees, logistic regression, random forest classifiers, and extreme gradient boosting 'XGBoost'[19] — on the whole training dataset (see *Overview*). We implemented a nested cross-validation strategy where the outer loop consisted of a 5-fold cross-validation and, within each inner loop, each fold's training set was used to tune that model's hyperparameters using a 3-fold cross-validation strategy. We applied this process 10 times, each time with the entire data sampled with replacement (bootstrap). We chose a 3-fold cross validation strategy for the inner loop, instead of the 5-fold used for the outer loop, due to the computational cost of running nested cross-validation on 12 thousand records.

Preventing data leakage

Unless otherwise noted, all cross-validation strategies used healthcare encounters (a.k.a, patient admissions), not individual reports. We split the reports this way to avoid "data leakage" — that is, having chest imaging reports from the same encounter appear in both the training and validation data. Thus, we ensured all reports from a given encounter can only be found on either the training or validation data (but not both). We also used nested cross-validation to prevent data leakage, as this avoids tuning hyperparameters on validation data.

Hyperparameter tuning

We used the Bayesian optimization implemented in *hyperopt* package (v.0.2.7)[20] for Python (v.3.10.12) to perform hyperparameter tuning. For each model trained, we performed cross-validation to obtain the mean log loss for each hyperparameter combination considered. We then selected the optimal combination of hyperparameters as the one yielding the lowest cross-validation mean log loss after at least 100 iterations. However, we employed an early-stopping rule to avoid running fruitless iterations.

Model comparison and feature importance

We used 95% confidence intervals to compare receiver operating characteristic (ROC) curves, areas under the ROC curves (AUROCs), and calibration curves across different models. For obtaining feature/token importance during training, we employed the default "importance" method that version 1.1.3 of the scikit-learn package implements for decision tree, logistic regression, and random forest algorithms, and version 1.7.4 of *xgboost* package for XGBoost. For decision trees and random forests, feature importance corresponds to the mean decrease in Gini impurity; for logistic regression, importances correspond to the mean value of coefficients in the fitted linear equation; and for XGBoost, the importance corresponds to the mean gain in predictive performance obtained by including a particular

feature in the trees. We additionally used Shapley-additive explanations (SHAP; v.0.42.0) values to obtain the feature/token importance on test sets.

Model testing for adjudication of bilateral infiltrates

Model selection resulted in XGBoost being chosen as best-performing algorithm for the task of adjudicating bilateral opacities in chest imaging reports. To assess the robustness of the XGBoost algorithm to different training sets, we tuned hyperparameters and then trained XGBoost models on all chest imaging reports from Hospital A (2013) and Hospital A (2016), separately. We then tested each of the two models on the two other chest imaging corpora *the model had not yet seen* by comparing the mean AUROC values after 5-fold cross validation. We used 10 resamples with replacement (bootstrap) to obtain 95% confidence intervals for the estimates of the mean AUROC values. This test also allowed us to scope out the generalizability of our approach.

However, we finally tested XGBoost's generalization by tuning its hyperparameters and then training it on the whole training set. We then applied this model to the 975 chest imaging reports from 100 MIMIC (2001-12) encounters, employing a similar cross-validation and bootstrapping approach as described above for testing XGBoost's robustness. It is this model which we carried forward for use in the pipeline, and we'll refer to this model as the "Bilateral Infiltrates Model".

We also evaluated the inter-rater disagreement rate for chest imaging reports from MIMIC (2001-12). For this purpose, we obtained two independent and blinded adjudications (one critical care physician and one internal medicine physician) for the 975 reports, and split imaging reports into three groups according to the Bilateral Infiltrates Model's output probabilities. For each group, we then calculated the fraction of imaging reports for which these independent raters disagreed on their labels.

<u>Adjudication of risk factors from physician notes</u>

Overview

We used 2034 labeled attending physician notes from Hospital A (2013) to develop our approaches to adjudicate ARDS risk factors and/or heart failure. However, not every one of the 2034 notes was labeled for each and every risk factor. For a full list of risk factors and heart failure language available for Hospital A (2013) notes, as well as how many notes were labeled for each, see SI: Regular expressions list 1 and Table S1. We did not have labeled attending physician notes available for Hospital System A (2016) or Hospital B (2017-2018). However, we had 887 labeled attending physician notes from MIMIC (2001-12) available, and we used them as our test set. Unlike notes from Hospital A (2013), each and every note from MIMIC (2001-12) was labeled by a critical care physician for the following criteria: pneumonia, aspiration, burns, pancreatitis, pulmonary contusion, sepsis, trauma, vasculitis, and cardiac surgery.

Capturing risk factors with regular expressions

The Berlin Definition of ARDS requires the presence of at least one risk factor — e.g., pneumonia, sepsis, shock, inhalation, pulmonary contusion, vasculitis, drowning, drug overdose — within seven days of non-cardiogenic acute respiratory failure. We preprocessed labeled attending physician notes from Hospital A (2013) to remove identifiable information from the text of these notes. We then used regular expressions (*regex* v2022.10.31) to match keywords related to risk factor and heart failure language that had labels available in Hospital A (2013) (see SI: Regular expression list 1, for a complete list of risk factors and regex patterns used to capture them). To validate this strategy, we ensured that this

regex approach captured 100% of the notes that had a positive label for a particular risk factor (or close to 100% as possible). We also corrected common spelling errors on important keywords, such as 'pneunonia', 'spetic' or 'cardigenic'.

Adjudication of risk factors with XGBoost - Feature engineering

Adjudicating the presence of some risk factors is not as simple as finding a particular keyword in a physician note. For example, a note stating "patient is unlikely to have pneumonia" should not be classified as evidence of pneumonia. To account for such possibilities, we tested an XGBoost approach for adjudicating those risk factors that had enough data for an ML technique (see below for our criteria for employing ML for a risk factor). We largely kept this approach similar to the adjudication of bilateral infiltrates from chest imaging reports, unless otherwise noted.We did not have a set of inclusion and exclusion language criteria to decide which words in an attending physician note are informative versus irrelevant. Therefore, we implemented a strategy in which, after matching a particular keyword through *regex*, we extract a text string from the note starting 100 characters prior to the occurrence of the keyword and extending 100 characters post the keyword. Subsequently, we tokenized and vectorized the 200-character strings as we did for the bilateral infiltrates' models — that is, created token frequency matrix for each 200-character string.

Adjudication of risk factors with XGBoost - Risk factor/model selection

Using the vectorized tokens, we trained XGBoost models for a select group of risk/heart failure factors. We only used a select group since not all risk factors were amenable to ML approaches: we chose risk factors that had more than 100 notes adjudicated in Hospital A (2013) notes corpus, and had relatively balanced yes/no proportions after regex-matching (between 33% and 66%, see Table S1). This resulted in the use of 1409 labeled attending notes from 337 patients for XGBoost model development. It also resulted in only trying an XGBoost approach for the following risk factors or heart failure language: pneumonia, aspiration, congestive heart failure, and sepsis. In other words, we did not use all 2034 labeled records from Hospital A (2013) to train each of the models since not every record had a label or a keyword for a given risk factor. For instance, only 636 notes in Hospital A (2013) included a label for pneumonia, whereas we were able to match 955 notes for pneumonia using regex. Therefore, our training dataset for each of the four models consisted of all notes that were regex-captured for that particular risk factor (e.g. 955 notes for the Pneumonia Model; see Table S1 for number of notes used to train each model). For notes that were regex-captured but did not have a label, we imputed the label as 'No', or zero (see Table S1 for label proportions after this imputation step).

We employed a similar nested cross-validation strategy as the one pursued for the adjudication of chest imaging reports. Notably, we also split the adjudicated notes into train and test sets by encounter, not note, to prevent data leakage. However, since we only had up to 1409 notes for training, this task was not as computationally expensive as it was for adjudicating chest imaging reports. Therefore, we used 100 resamples with replacement instead of 10, and always employed 5-fold cross validation for hyperparameter tuning.

Adjudication of risk factors with XGBoost – model testing

Model selection yielded pneumonia as the risk factor best suited to be adjudicated by an XGBoost approach. Therefore, we tested this "Pneumonia Model" by tuning XGBoost's hyperparameters and then training it on the 955 attending physician notes from Hospital A (2013) that were captured by the

pneumonia regex pattern. We then tested this Pneumonia Model on the 790 attending physician notes from 100 MIMIC (2001-12) that were captured by the same pneumonia regex pattern, employing a similar cross-validation and bootstrapping approach as described above for testing the Bilateral Infiltrates Model. It is this model which we carried forward for use in the pipeline.

Adjudication of risk factors with regular expressions

We only chose to use XGBoost models to adjudicate pneumonia, and therefore resorted to a regular expression approach for the rest of the risk factors and heart failure criteria. However, we employed two additional criteria to decide which of them would be adjudicated through regex: First, if a risk factor or heart failure criterion was labeled 'yes' in more than 80% of the captured notes, we adjudicated the risk factor or heart failure criterion with the same regex pattern developed for capturing it. Second, since many other risk factors and heart failure criteria had a more balanced proportion of yes/no labels after regex-matching, we refined the regex approach by implementing "exclusion" words or phrases for each risk factor or heart failure criterion (Regular expression list 2). The goal of this set of words was to reduce the false positives rate of risk factors or heart failure criteria as much as possible. We developed this set of exclusion words or phrases on notes from Hospital A (2013), and selected risk factors that were amenable to refinement with this approach.

Finally, we tested these two regex approaches on notes from MIMIC (2001-12) that were labeled for the same risk factors and heart failure criteria that were available for Hospital A (2013): aspiration, burns, pancreatitis, pulmonary contusion, sepsis, trauma, vasculitis. However, we only carried forward for use in the pipeline those regular expression patterns that fulfilled the above two regex criteria for any given hospital cohort.

<u>Design of ARDS adjudication pipeline: chaining Berlin Definition steps</u>

Each of the steps outlined above automates the adjudication of specific criteria in the Berlin Definition, with the modifications specified previously[7]. We integrate the different criteria into a single pipeline to build an automated ARDS adjudication pipeline.

The ARDS adjudication pipeline first flags encounters with at least one hypoxemia measurement (i.e., one instance of PF ratio ≤ 300 mm Hg while PEEP ≥ 5 cm H20), and then uses the predictions of the Bilateral Infiltrates Model to adjudicate presence of bilateral infiltrate language in chest imaging reports. Upon settling these two criteria, the pipeline flags whether the hypoxemia record and the report consistent with bilateral infiltrates have timestamps within 48 hours of each other (which we term "qualified hypoxemia"). In addition, at this step the pipeline ensures that the hypoxemia record was taken at or after intubation. Next, the pipeline uses the predictions of the Pneumonia Model to adjudicate pneumonia on all notes and uses regex to flag presence of other risk factors, heart failure language, and indicators of cardiogenic and noncardiogenic language. Specifically, for Hospital A (2013), we used a regex approach to adjudicate *sepsis, shock and its "cardiogenic" qualifier, inhalation, pulmonary contusion, burns, vasculitis, drowning, and overdose*. On the other hand, for MIMIC (2001-12), we used a regex approach to also adjudicate *aspiration and pancreatitis.* Finally, the pipeline flags whether an attending physician note has a timestamp that falls between one day prior to and seven days after the latter timestamp of any of the qualifying hypoxemia-bilateral infiltrates pairs. Once these annotations are integrated, the pipeline proceeds to adjudicate whether an ARDS diagnosis is warranted.

If any risk factor is identified in this time window for an encounter, regardless of whether *cardiogenic shock* is identified, the pipeline adjudicates the encounter as an ARDS case. If no risk factors

are identified, but only *cardiogenic shock* is identified, the pipeline adjudicates the encounter as a "No ARDS" case.

For all other encounters not meeting the risk factor or heart failure language criteria described above, the pipeline flags the case for objective heart failure assessment[7]. This assessment is done sequentially instead of by flagging. If any encounter had BNP greater than 100 pg/mL (an indicator of heart failure), the pipeline adjudicates "No ARDS" for such an encounter. The pipeline then considers the remaining encounters for each subsequent criteria, adjudicating "No ARDS" if the encounter had any of the following: left ventricular ejection fraction < 40%, cardiopulmonary bypass found in echocardiogram report, or at least two of the following present in the echocardiogram report: (i) left atrial diameter > 4 cm or left atrial volume index > 28 mL/m2, (ii) left ventricular hypertrophy, or (iii) Grade II or III diastolic dysfunction.

Any encounter that is not ruled out for ARDS after the objective heart failure assessment step is adjudicated as an ARDS case. That is, the pipeline adds these encounters to those adjudicated as positive for ARDS via risk factor identification.

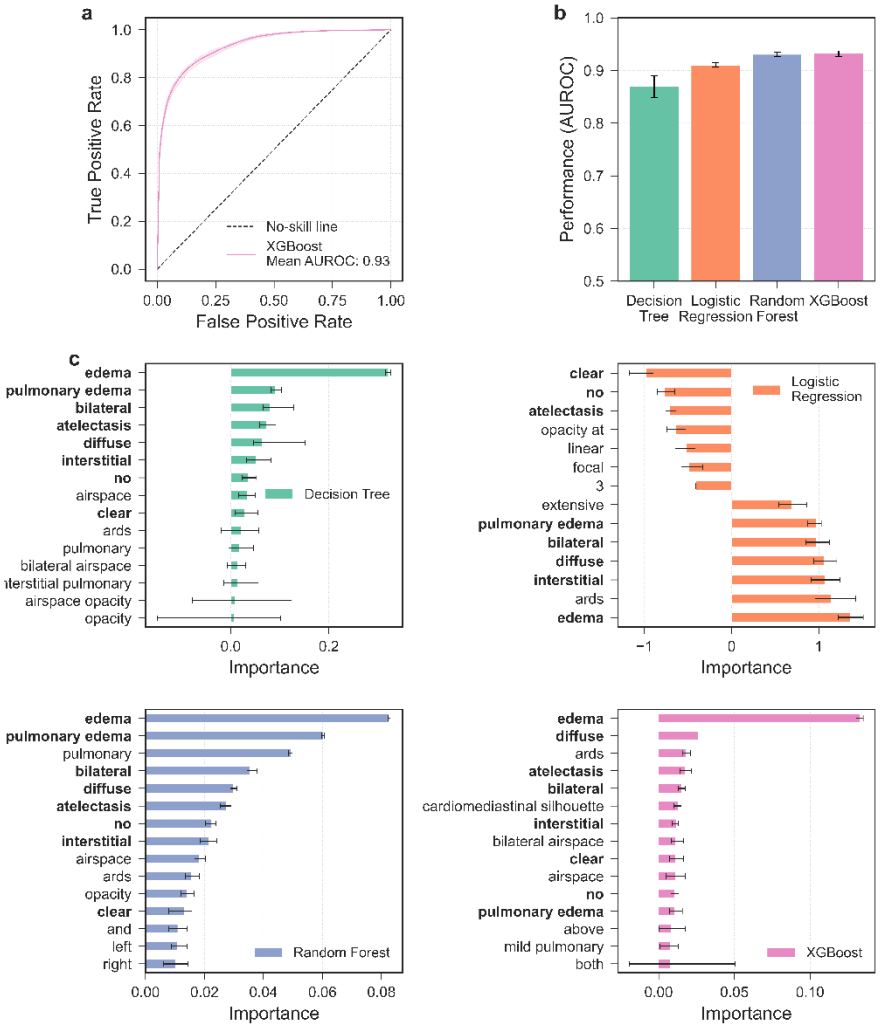
Results

Adjudication of bilateral infiltrates

Figure 1a shows the receiver operating characteristic (ROC) curves for the XGBoost model applied to chest imaging reports from the whole training set. We quantify model predictive performance using the areas under the ROC curves (AUROCs). We observe that once hyperparameters for each model are optimized, all models trained on chest imaging reports from the whole training set achieve AUROCs of at least 0.85 (Decision tree: AUROC = 0.87, 95%CI = [0.85, 0.89]; logistic regression: AUROC = 0.91, 95%CI = [0.91, 0.92]; random forest: AUROC = 0.93, 95%CI = [0.93, 0.93]; XGBoost: AUROC = 0.93, 95%CI = [0.93, 0.94]) (Fig. 1b).

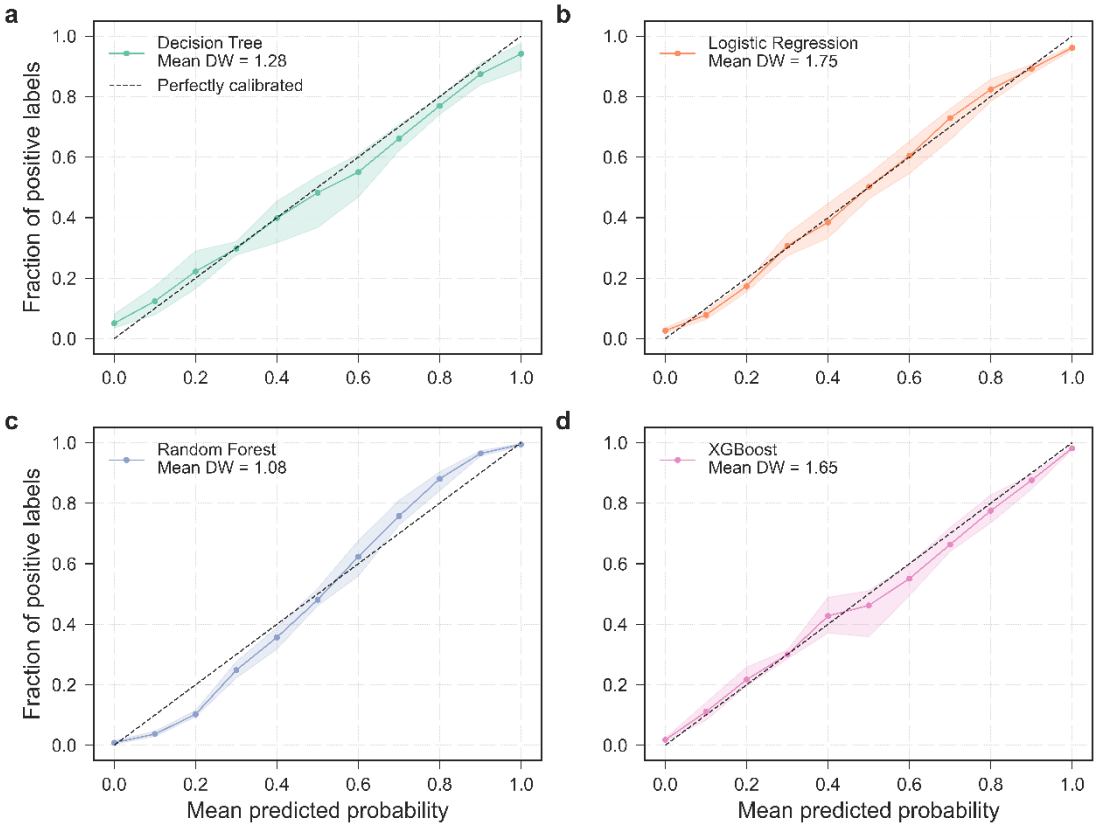
We next calculated the importance that each model assigned to the 200 tokens used as features. For decision trees and random forests, feature importance corresponds to the mean decrease in Gini impurity; for logistic regression, importances correspond to the mean value of coefficients in the fitted linear equation; and for XGBoost, the importance corresponds to the mean gain in predictive performance obtained by including a particular feature in the trees. Reassuringly, we find that the four models consistently identify tokens such as edema, bilateral, clear, and atelectasis as the most predictive (Fig. 1c). These tokens correspond closely to the inclusion/exclusion language we developed to address Berlin Definition shortcomings[7], which we also observed when implementing Shapley-additive explanations (SHAP) values to assess feature importance when applying the Bilateral Infiltrates Model (see methods, and below) on chest imaging reports from MIMIC (2001-12) (Fig. S1). For example, presence of the terms ‘bilateral’ or ‘edema’ has a positive impact on classification as positive for bilateral infiltrates. In contrast, presence of the terms ‘clear,’ ‘left’, or ‘right’ has a negative impact on classification as positive for bilateral infiltrates.

Figure 1. Machine learning (ML) models achieve high-performance in adjudicating the presence of bilateral infiltrates from chest imaging reports. Error bars show 95% confidence intervals for estimates of the mean obtained using bootstrapping. **a)** Receiver operating characteristic (ROC) curve for the XGBoost model trained on chest imaging reports from the whole training set. **b)** Bootstrapped mean area under the ROC (AUROC) show that all four ML approaches yield a predictive performance greater or equal to 0.85. **c)** Feature importances for the four different ML approaches considered. Features in bold are highly ranked in importance in all 4 approaches.



We then assessed how calibrated were the outcome probabilities from the models by comparing model outcome probability after training to actual probability of occurrence in considered encounters from the whole training set's chest imaging reports. Figure 2 suggests that logistic regression and XGBoost produce outcome probabilities that are well calibrated, which is expected given their use of similar loss functions for fitting (log-loss). In contrast, random forest produces poorly calibrated probabilities, being over-confident when forecasting with confidence levels lower than 50%, and under-confident with confidence levels greater than 50%. We thus select XGBoost as the implemented approach for our pipeline as it offers the highest predictive performance (AUROC = 0.93, 95%CI = [0.93, 0.94]) and well-calibrated forecasts on the whole training set (Durbin-Watson statistic = 1.65, 95%CI = [1.35-1.92]).

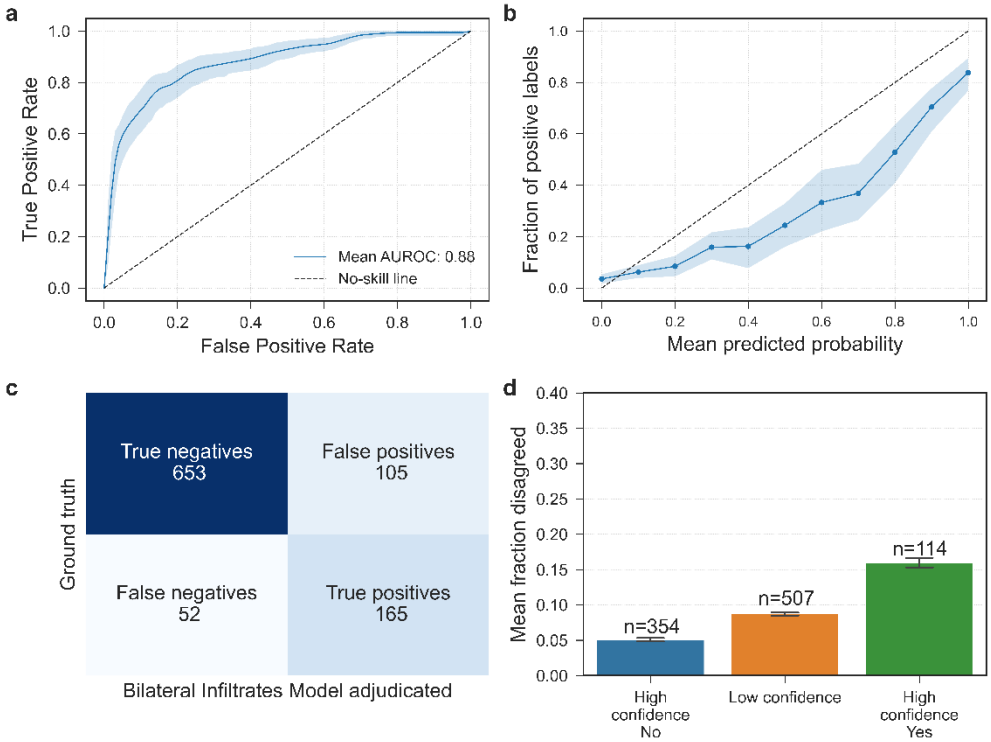
Figure 2. Calibration of produced probabilities in implemented ML models. We show calibration curves for models trained on the whole training set: **a)** decision tree, **b)** logistic regression, **c)** random forest, and **d)** XGBoost. A perfectly calibrated model would have a 1:1 relationship between fraction of positive labels and mean probabilities (i.e., it would overlay the diagonal line). The Durbin-Watson statistic, DW, probes for correlations in the residuals, if DW is close to 2, then one can rule out correlations in the residuals, implying good linear behavior.



To assess how an XGBoost model developed for a specific cohort generalizes to a different health system dataset, we tested the accuracy of an XGBoost model trained on Hospital A (2013 on chest imaging reports from Hospital B (2017-18) and MIMIC (2001-12) (Fig. S2a). We also trained a second XGBoost model on chest imaging reports from Hospital A (2016) and tested this model against chest imaging reports from Hospital B (2017-18) and MIMIC (2001-12) (Fig. S2b). In general, we found both XGBoost models performing at a mean AUROC of 0.85 or higher (lowest 95%CI of 0.84, highest 95%CI of 0.93). These results show that XGBoost could stably generalize well regardless of which available hospital system we chose to use for training.

We then tested the generalizability of our XGBoost approach by tuning hyperparameters and training XGBoost with the whole training dataset. We refer to this model as the Bilateral Infiltrates Model, and we tested it on 975 labeled chest imaging reports from MIMIC (2001-12). We reassuringly observed that the mean AUROC of this model reached 0.88 (95%CI = [0.85-0.91]), which is in line with the results of XGBoost models trained on specific cohorts (Fig. 3a). However, the calibration of the Bilateral Infiltrates Model on MIMIC (2001-12) suggested that the model is generally overconfident, except when estimated probabilities are less than 10% (Fig. 3b).

Figure 3. Evaluation of the bilateral infiltrates model on chest imaging reports from MIMIC (2001-12). Error bars show 95% confidence intervals for estimates of the mean obtained using bootstrapping. **a)** ROC curve for the bilateral infiltrates model tested on 975 bootstrapped chest imaging reports from MIMIC (2001-12). **b)** Calibration of probabilities by the bilateral infiltrates model when applied on 975 bootstrapped chest imaging reports from MIMIC (2001-12). **c)** Confusion matrix comparing MIMIC (2001-12) chest imaging report adjudications by the critical care physician (ground truth) against bilateral infiltrates model adjudications done at a 50% probability threshold. Notice that the numbers add up to 975. **d)** Comparing the disagreement rate between the critical care physician and the internal medicine physician against the confidence exhibited by the bilateral infiltrates model in adjudicating MIMIC (2001-12)'s chest imaging reports.



The calibration results prompted us to test whether the estimated probabilities by the Bilateral Infiltrates model could be associated with the inter-rater disagreement rates of the test set. To this end, we binned probabilities by the Bilateral Infiltrates Model into three groups: high-confidence 'No' (probability of bilateral infiltrates <10%), high-confidence 'Yes' (probability of bilateral infiltrates >90%), low confidence (all other probabilities of bilateral infiltrates). For each of these bins, we calculated the disagreement rate between the two independent rater — a critical care physician and an internal medicine physician. As seen on Figure 3d, the inter-rater disagreement was highest (16.0%) for the cases of high confidence 'Yes' predictions and lowest (5.1%) for the cases of high confidence "No". This, together with Figure 3b, suggests that the model is most reliable when human raters are most reliable as well, which happens when the model yields a very low probability of a chest imaging report showing bilateral infiltrates.

Finally, we evaluated how the BI model fared adjudicating bilateral infiltrates on individual chest imaging reports. We observed that, at the 50% probability threshold for binarizing probabilities into yes/no decisions, the bilateral infiltrates model exhibited a 76.0% sensitivity, and a 13.9% false positive rate in adjudicating bilateral infiltrates at the single report level (Fig. 3c). This results in an F1 score of 0.678 and an accuracy of 0.84. We also tried two different probability cutoffs aimed at maximizing either

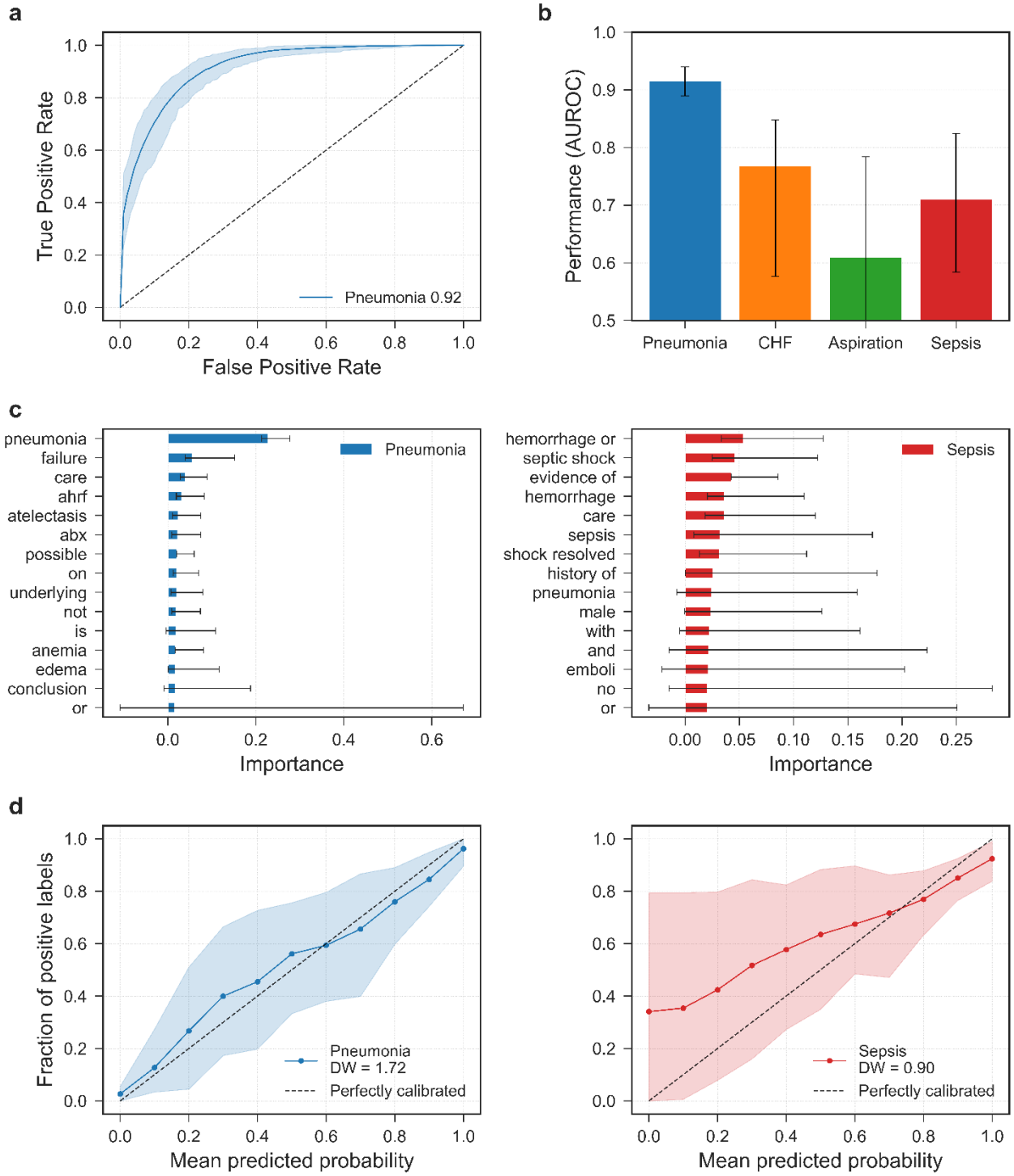
the F1 score (0.683, Fig. S3a), resulting in a 77.0% sensitivity and 13.9% false positive rate, or the accuracy (0.87, Fig. S3b), resulting in a 56.7% sensitivity and 4.0% false positive rate. Considering the problem of low ARDS recognition in clinical practice, our goal is to maximize sensitivity while preventing an excessive false positive rate. Therefore, the optimal F1 score probability cutoff appears as the choice that maximizes sensitivity for binarizing BI model's probabilities. We note that this optimal cutoff (49.6%) is within rounding error of the default cutoff of 50%.

Extracting ARDS risk factors in attending physician notes

We first developed regex patterns to match keywords for risk factors and heart failure language (see SI Regular expression list 1 for a full list of regular expressions). As shown in Figure S4, most of the developed regex patterns capture 100% of the notes that were labeled 'yes' for a particular risk factor in Hospital A (2013). When we count the total number of notes labeled as either yes or no, the most prevalent matches were sepsis (744 notes labeled vs. 748 notes regex-captured), pneumonia (636 notes labeled vs. 955 regex-captured), and shock (604 notes labeled vs. 607 regex-captured). For the heart failure criteria, the relevant matches were the cardiogenic keyword (176 notes labeled vs. 725 regex-captured) to qualify the matching of shock, and congestive heart failure (254 notes labeled vs. 352 regex-captured). We thus feel confident that the built regex patterns can match nearly the entirety of Hospital A (2013) notes labeled as "yes" for specific risk factors.

Next, we trained XGBoost models on attending physician notes from Hospital A (2013) cohort to adjudicate pneumonia, aspiration, congestive heart failure, and sepsis. We attempted a ML approach for these because at least 100 attending physician notes were labeled for them, and their labels have relatively balanced yes/no proportions after regex-matching (between 33% and 66%, with the exception of sepsis; see Table S1). We observed that the XGBoost model trained for pneumonia yielded the best discriminative performance during cross-validation (Pneumonia: AUROC = 0.92, 95%CI = [0.89, 0.94]; CHF: AUROC = 0.77, 95%CI = [0.58, 0.85]; Aspiration: AUROC = 0.61, 95%CI = [0.45, 0.78]; Sepsis: AUROC = 0.71, 95%CI = [0.58, 0.82]; Fig. 4a and4b), assigned the greatest and most stable importance to tokens that align with clinical expectations (Fig. 5c), and exhibited the best calibration (DW = 1.72, 95%CI = [1.31-2.28]; CHF: DW = 1.21, 95%CI = [0.41, 1.85]; Aspiration: DW = 0.61, 95%CI = [0.09, 1.35]; Sepsis: DW = 0.90, 95%CI = [0.30, 1.68]; Fig. 5d). These results show a stark contrast between the XGBoost models trained to adjudicate pneumonia and the XGBoost models trained to adjudicate congestive heart failure, aspiration, or sepsis. Thus, we decided against integrating these three XGBoost models into our pipeline, and only use the XGBoost approach to adjudicate pneumonia.

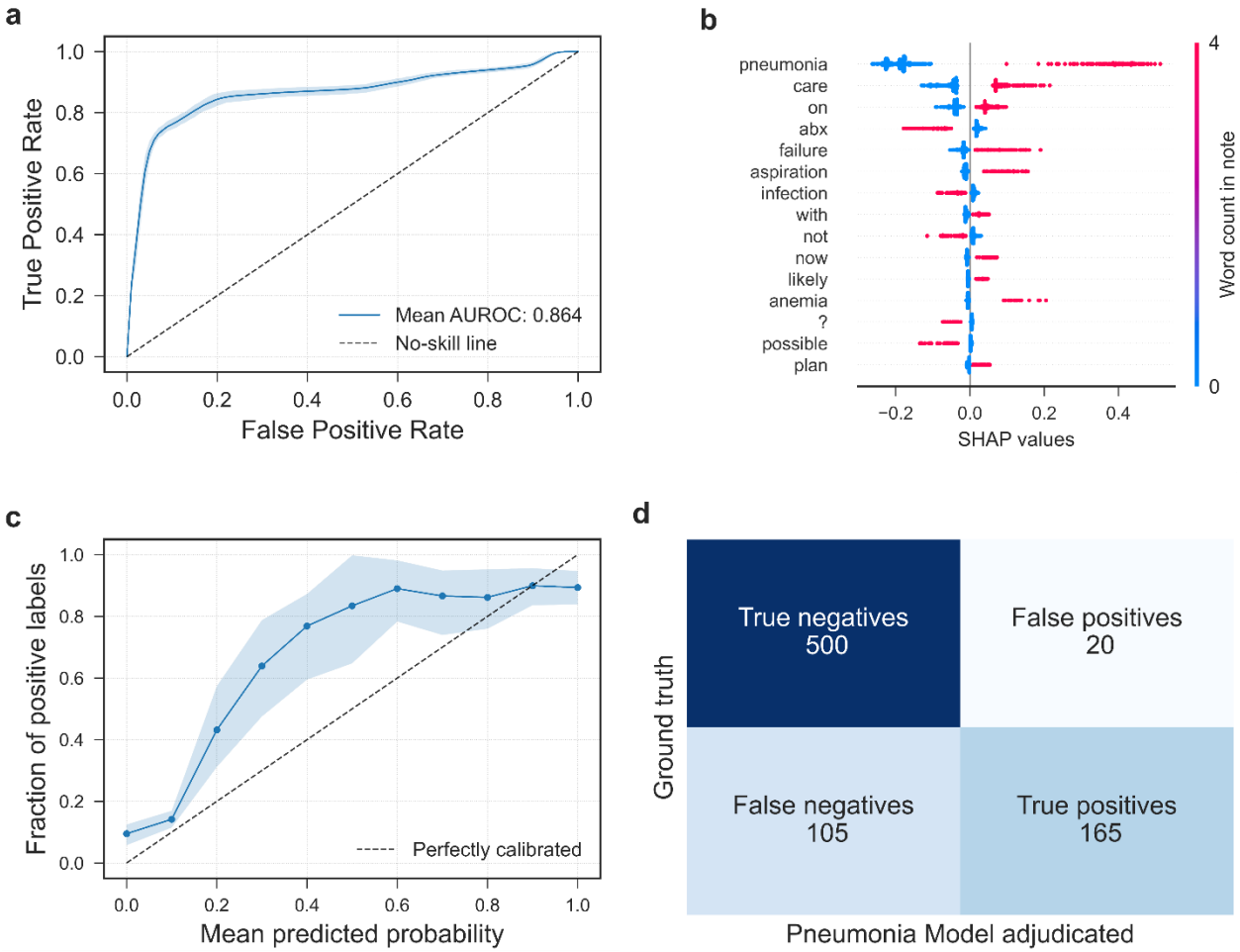
Figure 4. XGBoost model performance in adjudicating presence of some risk factors in attending physician notes. **a)** ROC curve showing the cross-validated performance of XGBoost model trained to adjudicate pneumonia on Hospital A (2013) attending physician notes. **b)** Bar plot showing cross-validated AUROCs of XGBoost models trained to adjudicate all the attempted risk factors, plus congestive heart failure, on Hospital A (2013) attending notes. **c)** Training set feature importances by the XGBoost models trained to adjudicate pneumonia and sepsis. **d)** Training set calibration curves for the XGBoost models trained to adjudicate pneumonia and sepsis.



We then tested this XGBoost approach for adjudicating pneumonia by tuning hyperparameters and training an XGBoost classifier on all of the 955 attending physician notes that were regex-captured for pneumonia. We call this model the Pneumonia Model, and we tested it on 790 attending physician notes from MIMIC (2001-12) that were regex-captured for pneumonia. We find that the mean AUROC for this model reached 0.86 (95%CI = [0.83-0.90]) (Fig. 5a), while making heavy use of the 'pneumonia' token for deciding the adjudication of pneumonia (Fig. 5b). However, the calibration of the Pneumonia Model on MIMIC (2001-12) suggested that the model is generally underconfident, except when estimated probabilities are greater than 80% (Fig. 5c). This last result is opposite to the calibration observed for the Bilateral Infiltrates Model.

Finally, we evaluated how the Pneumonia Model fared adjudicating pneumonia on individual attending physician notes. We observed that, at the 50% probability threshold for binarizing probabilities into yes/no decisions, the Pneumonia Model exhibited a 61.1% sensitivity and a 3.8% false positive rate in adjudicating pneumonia at the single note level (Fig. 5d). This results in an F1 score of 0.73 and an accuracy of 0.84. We also tried two different probability cutoffs aimed at maximizing either the F1 score (0.79, Fig. S5a), resulting in a 75.2% sensitivity and 7.7% false positive rate, or the accuracy (0.87, Fig. S5b), resulting in a 72.2% sensitivity and 5.8% false positive rate. Considering the problem of low ARDS recognition in clinical practice, our goal is to maximize sensitivity. Therefore, the 19.3% probability cutoff that maximizes the F1 score appears as the choice that maximizes sensitivity for binarizing Pneumonia Model's probabilities.

Figure 5. Evaluation of the Pneumonia Model on attending physician notes from MIMIC (2001-12). Error bars show 95% confidence intervals for estimates of the mean obtained using bootstrapping. **a)** ROC curve for the Pneumonia Model tested on 790 bootstrapped attending physician notes from MIMIC (2001-12) that were regex-captured for pneumonia. **b)** SHAP values for the top 15 words in terms of their impact on Pneumonia Model's estimated probabilities. **c)** Calibration of probabilities by the Pneumonia Model when applied on 790 bootstrapped attending physician notes from MIMIC (2001-12). **d)** Confusion matrix comparing MIMIC (2001-12) attending physician notes pneumonia labels by the critical care physician (ground truth) against Pneumonia Model adjudications done at a 50% probability threshold. Notice that the numbers add up to 790.

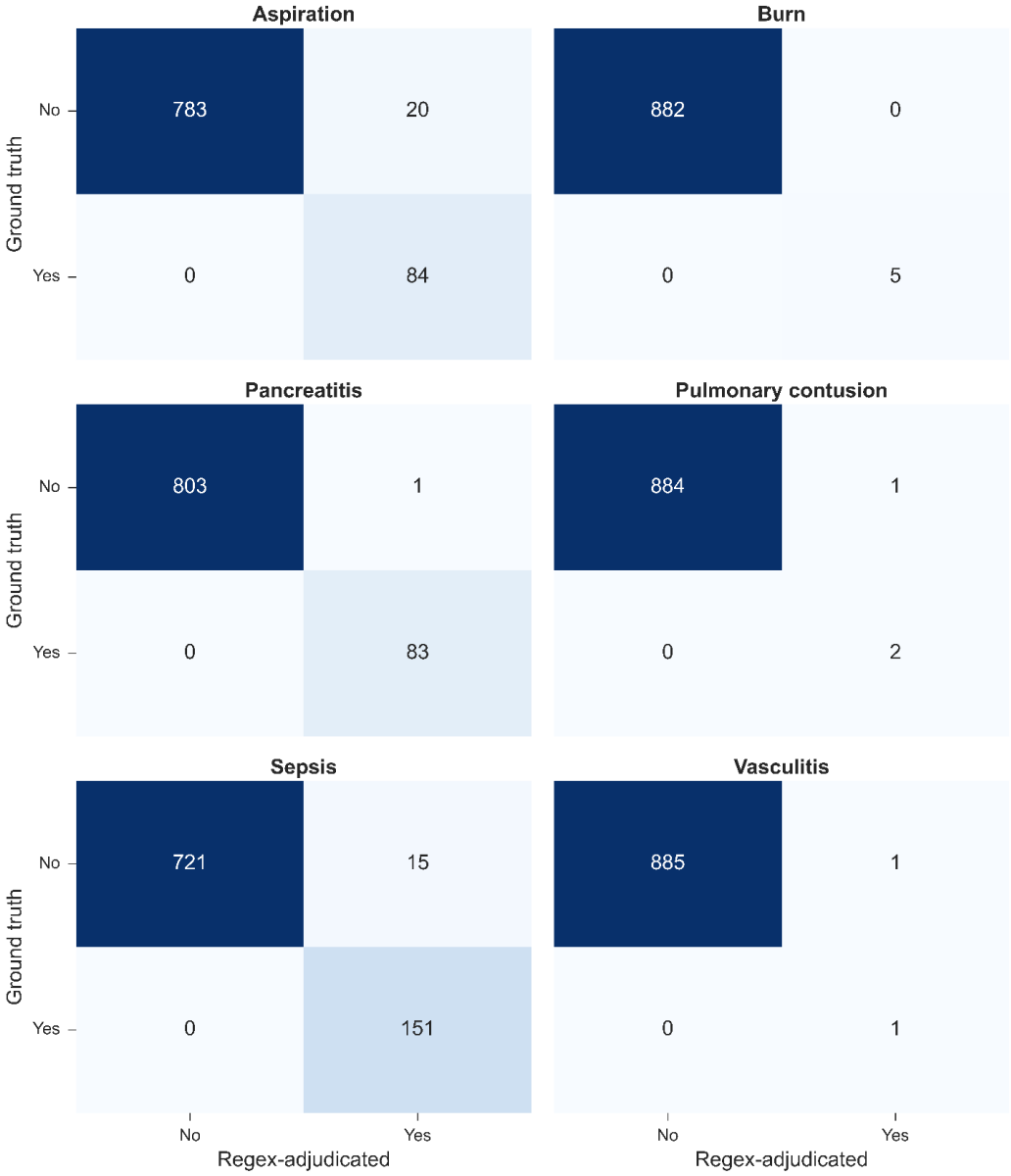


For all other risk factors and heart failure criteria, we resorted to regex adjudication through two approaches: (1) If a risk factor or heart failure criterion was labeled 'yes' in more than 80% of the captured notes, we adjudicated the risk factor using the same regex pattern created for capturing it. These included shock, cardiogenic, inhalation, and pulmonary contusion, vasculitis, drowning, and overdose (Table S1 and Fig. S4). (2) We reduced the number of false positives of some risk factors or heart failure criteria by developing a set of "exclusion" words or phrases and applying them after regex-capture (Regular expression list 3). We implemented this second approach for sepsis and shock, and observed a much-reduced false positive rate when compared to only using the regex patterns designed for capturing them (Fig. S6).

We then applied these regex approaches to adjudicate risk factors or heart failure criteria on attending physician notes from MIMIC (2001-12). Specifically, we adjudicated risk factors for which we had labels available to directly test the patterns developed for Hospital A (2013). Reassuringly, the regex patterns exhibited good test characteristics, most notably for burns, pancreatitis, sepsis, and aspiration (Fig. 6).

Based on the overall results, we incorporated *pneumonia, sepsis, shock and its "cardiogenic" qualifier, aspiration, inhalation, pulmonary contusion, pancreatitis, burns, vasculitis, drowning, and overdose* into the pipeline to adjudicate risk factors or heart failure language.

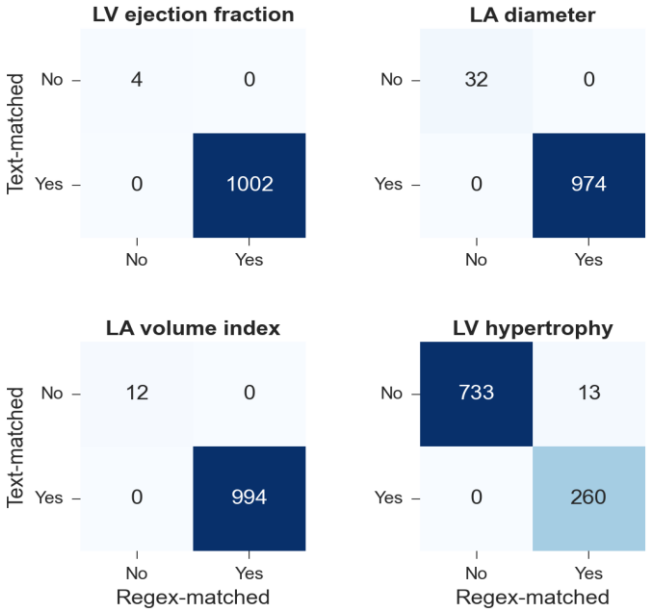
Figure 6. Risk factor adjudication performance of regular expressions on MIMIC (2001-12). All regular expressions developed adjudicate 100% of the attending physician notes from MIMIC (2001-12) labeled as "yes" for each risk factor. In terms of precision, burns, pancreatitis, sepsis, and aspiration exceeded 80%.



Adjudication of heart failure from echocardiogram (echo) reports

The criteria for the objective assessment of heart failure rely on the following six factors: left ventricular ejection fraction, cardiopulmonary bypass, left atrial diameter, left atrial volume index, left ventricular hypertrophy, and grade II or III diastolic dysfunction. Because echo reports are highly standardized, it is possible to extract these factors from the reports using regex. Moreover, we had access to echo reports from Hospital A (2013) which were previously text-matched, enabling us to validate our regex approach.

Figure 7. Confusion matrices comparing the flagging performance regex-matching against text-matching for 'Left ventricular ejection fraction', 'Left atrial dimension/diameter', 'Left atrial volume index', and 'Left ventricular hypertrophy'. Note the large discrepancy for the annotations of 'left ventricular hypertrophy', which is explained in text.



Using the regex patterns listed in the SI, we analyze Hospital A (2013)'s echo reports for the presence or absence of each of the six factors of interest. Figure 7 demonstrates that not all six factors were present in every echo report. For three of the six factors — 'left ventricular ejection fraction', 'left atrial dimension/diameter', and 'left atrial volume index'— we found excellent agreement between regex and text-matching. Two of the other three, 'cardiopulmonary bypass' and 'diastolic function', were not text-matched, so no comparison can be made. For 'left ventricular hypertrophy', the regex-matching procedure correctly captured the desired language, indicating that the original text-matching procedure failed to identify 13 echo reports. In addition, we validated the numerical values extracted through this regex approach by randomly selecting 10% of echo reports for visual inspection of values and comparing against values extracted through regex. We found 100% concordance between values extracted and those retrieved manually (Table S2).

Adjudication of ARDS for Hospital A (2013) cohort

We are now ready to compare the performance of our complete pipeline against the previously reported ARDS adjudication[7]. We use a threshold of 50% to map estimated probabilities into binary yes/no decisions for both the Bilateral Infiltrates Model and the Pneumonia Model.

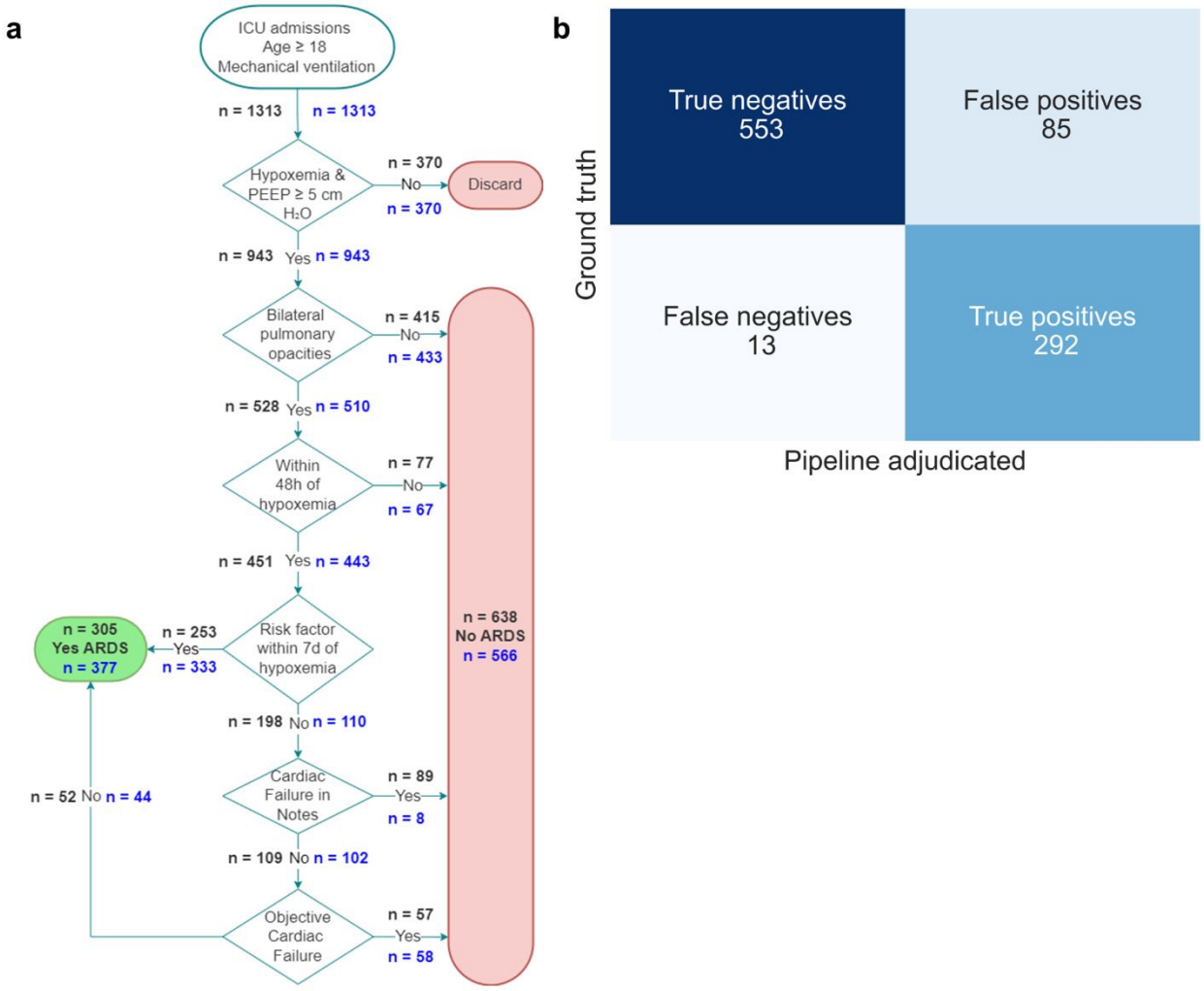
We conduct the evaluation of our pipeline for the 943 patients in the Hospital A (2013) ARDS adjudication cohort who were 18 years and older, received invasive mechanical ventilation, and had acute hypoxemic respiratory failure (Fig. 8). 143 patients had no chest imaging report available and were adjudicated as negative for ARDS. The remaining 800 patients had at least one chest imaging report available. The Bilateral Infiltrates Model adjudicated bilateral infiltrates within 48 h of a hypoxemic episode for 510 patients. Of these 510 patients, 443 had at least one of the qualified hypoxemic events occurring post-intubation and 333 had a risk factor within 7 days of the qualified hypoxemia event and were adjudicated as being positive for ARDS.

The remaining 110 patients were then evaluated for heart failure. For 8 patients, the physician notes indicated heart failure and they were adjudicated as negative for ARDS. The last 102 patients were then adjudicated using the objective heart failure assessment step; 58 were adjudicated to have heart failure and thus negative for ARDS and the remaining 44 were adjudicated as positive for ARDS. In total, the pipeline adjudicated 377 patients as positive for ARDS and 566 as negative for ARDS.

In summary, using a simple 50% probability cutoff for both ML algorithms, our pipeline yields close agreement with the physician adjudication of ARDS for this cohort[7] (Fig. 8a). Specifically, the pipeline yields a sensitivity or true positive rate of 95.7% on this cohort, which compares most favorably to the 19% ARDS diagnosis rate we previously found on this cohort[7]. Importantly, this high sensitivity is achieved while maintaining a low 13.3% rate of false positives.

We also explored different probability threshold choices for the Bilateral Infiltrates Model, and found a sensitivity of 97.0% with a false positive rate of 16.0% when using an optimal F1 score threshold, and a sensitivity of 96.4% with a false positive rate of 14.6% when using an optimal accuracy threshold (Fig. S7a).

Figure 8. Machine learning computational pipeline for adjudication of Hospital A (2013) cohort yields a small fraction of false negatives and an acceptable fraction of false positives. a) Flowchart of ARDS diagnosis by computational pipeline (blue) vs. physician from previous publication[7] (black). **b)** Confusion matrix comparing physician adjudication from previous publication[7] against ML computational adjudication pipeline.

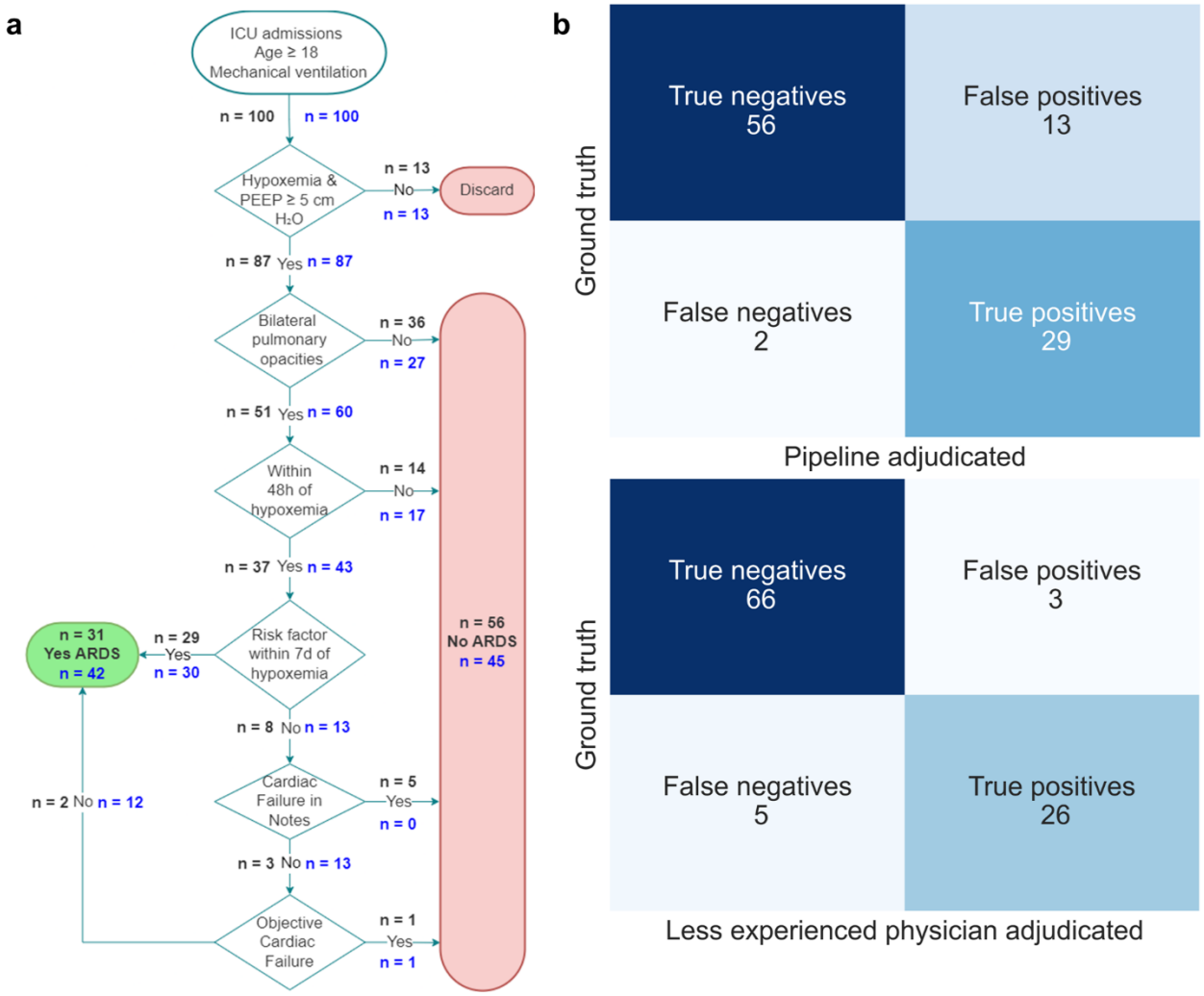


Adjudication of ARDS for MIMIC (2001-12) labeled subset

We applied our automated ARDS adjudication to the 100 patient encounters in MIMIC (2001-12), and then compared a critical care physician's adjudication against the pipeline's (Fig. 9). As discussed previously, we chose probability cutoffs that maximized the F1 score since these improved the sensitivity of the models on MIMIC (2001-12) data. Reassuringly, we find that the overall performance of our pipeline on MIMIC (2001-12) is strikingly similar to its performance on the development cohort[7] (Fig. 9b). Specifically, the pipeline yields a sensitivity or true positive rate of 93.5% on this cohort, which compares favorably to the 22.6% ARDS documentation rate we found in this subset. This high sensitivity is achieved while maintaining a relatively low 18.8% false positive rate. Moreover, the false negative rate of the pipeline adjudication (6.9%) is lower than that of a physician not trained in critical care medicine (16.1%), highlighting the pipeline's potential to aid physicians in ARDS diagnosis.

We also explored different probability threshold choices for both the Bilateral Infiltrates Model and the Pneumonia Model, and found a sensitivity of 93.5% with a false positive rate of 18.8% when using the default 50% probability threshold for both models, and a sensitivity of 96.4% with a false positive rate of 14.6% when using an optimal accuracy threshold for both models (Fig. S6b). We note that metrics from the default threshold and the optimal F1 score threshold are identical due to the similar cutoffs from both approaches when binarizing probabilities by the Bilateral Infiltrates Model.

Figure 9. Machine learning computational pipeline for adjudication of MIMIC (2001-12) cohort yields a small fraction of false negatives and a manageable fraction of false positives. a) Flowchart of ARDS diagnosis by computational pipeline (blue) vs. critical care physician (black). **b)** Confusion matrix comparing critical care physician adjudication (ground truth) against ML computational adjudication pipeline (top panel), and critical care physician adjudication (ground truth) against a less experienced physician adjudication.



Discussion

We believe that computational pipelines aiming to help physicians with the diagnosis of complex conditions must follow two principles. First, they should act as physician aids, not physician replacements. That is, they should flag a potential diagnosis for consideration by the responsible physician, rather than mandate a diagnosis as certain. Second, and a consequence of the first, they should provide interpretable insights. Others have pointed out[21] that machine learning (ML) should only be considered as the final decision maker for problems that can be interpreted as deterministic, such as differentiating a dog from a cat in a photo. However, for tasks where the characteristics of the two classes overlap and the outcome of the decision has important consequences, such as medical decision-making, the ML approaches should be used to provide an estimation of probabilities, not a final determination.

In this study, we construct and validate a ML pipeline for automating the adjudication of ARDS according to the Berlin Definition based on data from EHRs. We constructed high performing decision tree-based models (XGBoost) to adjudicate chest imaging reports for bilateral infiltrate language and attending physician notes for the presence of pneumonia. While large language models (LLMs) are growing in popularity and are, at times, seen as a panacea, we believe that in this context, using an interpretable model that can be easily versioned offers benefits that would be lost even if LLMs could produce higher accuracy. The tree-based methods we consider estimate probabilities, the first stage of any classification problem, and enable physicians to optimize the false positive vs. false negative tradeoff by adjusting the decision cutoff. In addition, our implementation of XGBoost allows language-level interpretation of estimated probabilities, which can enhance physician trust in ML models.

Supplementing the above XGBoost models with regular expressions to identify other ARDS risk factors and heart failure, and additional structured data to objectively rule out heart failure,[7] enabled the automated adjudication of the complete set of Berlin Definition criteria. This pipeline demonstrated excellent test characteristics, including false negative and false positive rates of 6.6% and 12.5%, respectively, at the 50% decision cutoff. Significantly, our validation of the generalizability of the pipeline to a subset of the MIMIC-III dataset yielded similarly high-performance levels. We firmly believe that false positive and false negative rates are preferred metrics for evaluation of performance in this context. The primary challenge faced in our study's clinical setting is ARDS under-diagnosis, so a very low false negative rate is extremely important. An important second challenge is decision-making fatigue, so a false positive rate that does not create an excess of unnecessary decisions will result in greater adoption. Other clinical contexts may demand more attention to the false positive rate or other performance metrics; our methods are fully adaptable to other potential clinical problem.

To our knowledge, this study is the first to automate the entire Berlin Definition process using ML and rules-based methods in a multi-center, open-source, generalizable manner. Previous attempts at automated ARDS adjudication used single-center EHR data to adjudicate individual Berlin criteria or used non-reproducible methods. Afshar et. al. used text features in chest imaging reports for ARDS identification, achieving a maximum AUROC of 0.80 for that task[22]. However, our work identifies ARDS by considering data beyond chest imaging reports. Sathe et. al. developed EHR-Berlin, evaluating the Berlin Definition using ML and rules-based methods, but their focus was limited to COVID-19 patients[15]; by using a cohort of patients who were already defined as having an ARDS risk factor, they effectively eschewed the need to identify ARDS risk factors or heart failure. In contrast, our study considers any adult patient placed on mechanical ventilation, which requires evaluating all Berlin Definition components. Finally, Song and Li developed a fully rules-based tool that automates the entire Berlin Definition, both achieving identical high performance[13,14]. However, their models were constructed within

a single hospital over a single time period, which may not be generalizable across different healthcare systems.

Machine learning offers a powerful solution for efficiently analyzing large volumes of data that would otherwise require countless hours of human effort. A recent study that combined NLP techniques with manual chart abstraction to evaluate clinical trial outcomes confirms this assertion. The time spent on retrospective review decreased from 2,000 costly and laborious hours for manual chart abstraction to just 34.3 hours with NLP-screened manual chart abstraction[23]. Similarly, our pipeline demonstrates the ability to adjudicate ARDS for multiple cohorts of hundreds of patients in under five minutes by training XGBoost models at runtime (with even faster runtimes possible using pre-trained models on inference mode).

A potential limitation is that previous studies have shown significant variability in the diagnosis of ARDS *among critical care physicians*, particularly in relation to interpretation of chest imaging[18]. We attempted to mitigate concerns raised by this challenge by only relying on chest imaging reports written by radiologists. This allowed us to use previously developed Berlin Definition-based inclusion and exclusion language as a guideline for critical care physicians reviewing chest imaging reports to minimize interrater disagreements[7]. As a consequence, we are also able to leverage this inclusion and exclusion language for the NLP processing of chest imaging reports needed for our ML approach. We believe this choice allowed us to increase the signal-to-noise ratio of the data used for ML development; however, we recognize that this choice might create additional challenges if implementing this pipeline for real-time use.

We also explored the relationship between interrater disagreement and ML model confidence. We found that lower disagreement rates among our raters correlated with relatively lower model confidence of "yes", indicating that our algorithm can confidently "discard" cases which are not likely to have bilateral infiltrates. Interestingly, we also observed that the human raters exhibited higher levels of disagreement when the model had high model confidence of "yes". We speculate this could be attributed to MIMIC (2001-12) containing a greater proportion of reports that are not consistent with bilateral infiltrates (78%).

Another limitation of our approach concerns the adjudication of ARDS risk factors. We faced significant challenge in implementing machine learning techniques for parsing attending physician notes due to the lack of clearly defined inclusion/exclusion language for adjudicating ARDS risk factors. In addition, we had only 744 attending physician notes labeled for sepsis, the most frequent of the risk factors for ARDS, compared to more than 12 thousand labeled chest imaging reports. Fortunately, in practice it turns out that even a standard regex approach yielded high overall performance and a small fraction of false positives.

A possible path to improve performance could involve refining the note-parsing process using advanced natural language processing techniques such as Med-BERT, cTAKES, or leveraging the capabilities of open-source LLMs. The latter technique has the potential to eliminate laborious pre-processing steps and facilitate the development of general-purpose models instead of task-specific ones[23]. While such approaches offer exciting opportunities, we must remain cautious as it is in the interest of patients and physicians to implement approaches that prioritize interpretability and transparency. Not to mention the cost-effectiveness of developing and deploying pipelines such as ours instead of those relying on large language models.

Another limitation of our study is the relatively small number of medical centers and time periods for which we were able to obtain data. However, their number is significantly larger than what is typical

in the literature. While concerns about the lack of robustness of our approach are largely answered by the high performance obtained using different models and different data sets for adjudicating bilateral infiltrates, the same is not true to the same extent for determination of risk factors or for the adjudication of echocardiogram reports. It is possible that implementation of our pipeline at additional medical centers would require the development of new regex patterns. However, the information required by the pipeline is quite standardized, assuaging some of those concerns.

Our pipeline compellingly answers the specific question being posed: can we automate the identification of ARDS in a way that is clinically relevant? In the clinical realm, minimizing the false negative rate (at the expense of a still manageable but higher false positive rate) means applying ARDS treatment to patients who do not have ARDS, which is likely to be less harmful than not treating patients who do have ARDS but were not recognized as such [26,27]. The pipeline powerfully addresses this clinical goal.

While our pipeline will aid us and quality reviewers at our institutions during retrospective reviews, its greatest potential impact lies in its integration with clinical decision support systems, enabling timely alerts to critical care physicians about the probability of ARDS in their patients. Future studies should evaluate the pipeline in several ways, such as a physician fully trusting the pipeline when ML exhibits high confidence in its probability estimations, or a physician double-checking a case only if their adjudication differs from the pipeline.

To properly implement this pipeline, decision theory considerations should be addressed. The XGBoost models in our pipeline generate probabilities that need to be binarized into “yes” or “no” for each component of the Berlin Definition. In this regard, Youden's J statistic is used by some to identify optimal thresholds that yield a good balance between false negatives and false positives[24]. While the use of Youden's J for cutoff determination is common practice in theoretical studies, it assumes a similar degree of undesirability for false positives and false negatives while implicitly using disease prevalence as a cost ratio[25]. We believe that implementation is more likely to be successful if a health system explicitly considers the particulars of ARDS when deciding on optimal probability cutoffs. Given the poor recognition of ARDS in clinical practice, prioritizing comparatively low false negative rates is crucial for making life-saving decisions, such as implementing low tidal volume ventilation and prone-positioning strategies[28]. This benefit must be balanced with the potential risk of alert fatigue caused by excessive false positives[29–31]. An implementation study that explores the attitudes of critical care physicians towards the balance of false positives and false negatives could provide valuable insights for implementing decision-support tools like our pipeline.

References

1. James, J. T. A New, Evidence-based Estimate of Patient Harms Associated with Hospital Care. *J. Patient Saf.* **9**, 122–128 (2013).
2. Makary, M. A. & Daniel, M. Medical error—the third leading cause of death in the US. *BMJ* i2139 (2016) doi:10.1136/bmj.i2139.
3. Landrigan, C. P. *et al.* Temporal Trends in Rates of Patient Harm Resulting from Medical Care. *N. Engl. J. Med.* **363**, 2124–2134 (2010).
4. DeGrave, A. J., Janizek, J. D. & Lee, S.-I. AI for radiographic COVID-19 detection selects shortcuts over signal. *Nat. Mach. Intell.* **3**, 610–619 (2021).
5. The ARDS Definition Task Force*. Acute Respiratory Distress Syndrome: The Berlin Definition. *JAMA* **307**, 2526–2533 (2012).
6. Bellani, G. *et al.* Epidemiology, Patterns of Care, and Mortality for Patients With Acute Respiratory Distress Syndrome in Intensive Care Units in 50 Countries. *JAMA* **315**, 788–800 (2016).
7. Weiss, C. H. *et al.* Low Tidal Volume Ventilation Use in Acute Respiratory Distress Syndrome*. *Crit. Care Med.* **44**, (2016).
8. Weiss, C. H. *et al.* A Critical Care Clinician Survey Comparing Attitudes and Perceived Barriers to Low Tidal Volume Ventilation with Actual Practice. *Ann. Am. Thorac. Soc.* **14**, 1682–1689 (2017).
9. Koenig, H. C. *et al.* Performance of an automated electronic acute lung injury screening system in intensive care unit patients*. *Crit. Care Med.* **39**, (2011).
10. Herasevich, V., Yilmaz, M., Khan, H., Hubmayr, R. D. & Gajic, O. Validation of an electronic surveillance system for acute lung injury. *Intensive Care Med.* **35**, 1018–1023 (2009).
11. Laffey, J. G., Pham, T. & Bellani, G. Continued under-recognition of acute respiratory distress syndrome after the Berlin definition: what is the solution? *Curr. Opin. Crit. Care* **23**, (2017).
12. Zeiberg, D. *et al.* Machine learning for patient risk stratification for acute respiratory distress syndrome. *PLOS ONE* **14**, e0214465 (2019).
13. Song, X., Weister, T. J., Dong, Y., Kashani, K. B. & Kashyap, R. Derivation and Validation of an Automated Search Strategy to Retrospectively Identify Acute Respiratory Distress Patients Per Berlin Definition. *Front. Med.* **8**, (2021).
14. Li, H. *et al.* Rule-Based Cohort Definitions for Acute Respiratory Distress Syndrome: A Computable Phenotyping Strategy Based on the Berlin Definition. *Crit. Care Explor.* **3**, e0451 (2021).
15. Sathe, N. A. *et al.* Evaluating construct validity of computable acute respiratory distress syndrome definitions in adults hospitalized with COVID-19: an electronic health records based approach. *BMC Pulm. Med.* **23**, 292 (2023).
16. Johnson, A. E. W. *et al.* MIMIC-III, a freely accessible critical care database. *Sci. Data* **3**, 160035 (2016).
17. Johnson, A., Pollard, T. & Mark, R. MIMIC-III Clinical Database. [object Object] https://doi.org/10.13026/C2XW26 (2015).
18. Sjoding, M. W. *et al.* Interobserver Reliability of the Berlin ARDS Definition and Strategies to Improve the Reliability of ARDS Diagnosis. *Chest* **153**, 361–367 (2018).
19. Chen, T. & Guestrin, C. XGBoost: A Scalable Tree Boosting System. in *Proceedings of the*

22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining 785–794 (Association for Computing Machinery, New York, NY, USA, 2016). doi:10.1145/2939672.2939785.
20. Bergstra, J., Yamins, D. & Cox, D. D. Making a Science of Model Search: Hyperparameter Optimization in Hundreds of Dimensions for Vision Architectures. in *Proceedings of the 30th International Conference on International Conference on Machine Learning - Volume 28* I-115-I–123 (JMLR.org, 2013).
21. van den Goorbergh, R., van Smeden, M., Timmerman, D. & Van Calster, B. The harm of class imbalance corrections for risk prediction models: illustration and simulation using logistic regression. *J. Am. Med. Inform. Assoc.* **29**, 1525–1534 (2022).
22. Afshar, M. *et al.* A Computable Phenotype for Acute Respiratory Distress Syndrome Using Natural Language Processing and Machine Learning. *AMIA Annu. Symp. Proc. AMIA Symp.* **2018**, 157–165 (2018).
23. Lee, R. Y. *et al.* Assessment of Natural Language Processing of Electronic Health Records to Measure Goals-of-Care Discussions as a Clinical Trial Outcome. *JAMA Netw. Open* **6**, e231204–e231204 (2023).
24. Youden, W. J. Index for rating diagnostic tests. *Cancer* **3**, 32–35 (1950).
25. Smits, N. A note on Youden's Jand its cost ratio. *BMC Med. Res. Methodol.* **10**, 89 (2010).
26. Serpa Neto, A. *et al.* Association Between Use of Lung-Protective Ventilation With Lower Tidal Volumes and Clinical Outcomes Among Patients Without Acute Respiratory Distress Syndrome: A Meta-analysis. *JAMA* **308**, 1651–1659 (2012).
27. Determann, R. M. *et al.* Ventilation with lower tidal volumes as compared with conventional tidal volumes for patients without acute lung injury: a preventive randomized controlled trial. *Crit. Care* **14**, R1 (2010).
28. Petrucci, N. & De Feo, C. Lung protective ventilation strategy for the acute respiratory distress syndrome. *Cochrane Database Syst. Rev.* (2013) doi:10.1002/14651858.CD003844.pub4.
29. Ancker, J. S. *et al.* Effects of workload, work complexity, and repeated alerts on alert fatigue in a clinical decision support system. *BMC Med. Inform. Decis. Mak.* **17**, 36 (2017).
30. Lee, E. K., Wu, T.-L., Senior, T. & Jose, J. Medical Alert Management: A Real-Time Adaptive Decision Support Tool to Reduce Alert Fatigue. *AMIA. Annu. Symp. Proc.* **2014**, 845–854 (2014).
31. Cvach Maria. Monitor Alarm Fatigue: An Integrative Review. *Biomed. Instrum. Technol.* **46**, 268–277 (2012).

Acknowledgements: The authors thank Catherine Gao for insightful discussions and suggestions.

Funding: FX was supported in part by the National Institutes of Health Training Grant (T32GM008449) through Northwestern University's Biotechnology Training Program; R.A.K.R. CHW was supported by the National Heart Lung and Blood Institute (R01HL140362 and K23HL118139). LANA was supported by the National Heart Lung and Blood Institute (R01HL140362). LANA and FX are supported by the National Institute of Allergy and Infectious Diseases (U19AI135964).

Author Contributions: **FM -** Methodology, Software, Validation, Data Curation, Writing – Original Draft, Writing – Review & Editing, Visualization. **HAL -** Methodology, Software, Validation, Data Curation, Writing – Review & Editing. **HTN -** Software, Validation, Data Curation. **MB -** Methodology, Validation, Data Curation, Writing – Review & Editing. **FX -** Methodology, Software, Validation, Data Curation, Writing – Review & Editing, Visualization. **JK -** Data Curation. **ELC -** Data Curation, Writing – Review & Editing. **CHW -** Conceptualization, Methodology, Validation, Data Curation, Resources, Writing – Original Draft, Writing – Review & Editing, Visualization, Supervision, Project Administration, Funding Acquisition. **LANA -** Conceptualization, Methodology, Software, Validation, Formal Analysis, Resources, Writing – Original Draft, Writing – Review & Editing, Visualization, Supervision, Project Administration, Funding Acquisition.

Competing Interests: The authors declare that they have no competing interests.

Data Availability. The datasets analyzed in this study will be made available upon publication at ARCH repository hosted by Northwestern University (https://arch.library.northwestern.edu).

Code Availability. The Python code to reproduce the reported results will be made available upon publication at the Amaral lab GitHub repository (https://github.com/amarallab).