

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

3

4 Félix L. Morales<sup>1†</sup>, Feihong Xu<sup>2</sup>, Hyojun Ada Lee<sup>1</sup>, Helio Tejedor Navarro<sup>1,3</sup>, Meagan A. Bechel<sup>4,5</sup>, Eryn

5 L. Cameron<sup>6</sup>, Jesse Kelso<sup>6</sup>, Curtis H. Weiss<sup>1,7</sup>, Luís A. Nunes Amaral<sup>1,3,8,9,10\*</sup>

6

7 [1] Department of Chemical and Biological Engineering, Northwestern University, Evanston, IL

8 [2] Interdepartmental Biological Sciences Program, Northwestern University, Evanston, IL

9 [3] Northwestern Institute on Complex Systems, Northwestern University, Evanston, IL

10 [4] Medical Scientist Training Program, Northwestern University Feinberg School of Medicine, Chicago, IL

11 [5] Department of Radiology, Emory University, Atlanta, GA

12 [6] Department of Medicine, Endeavor Health, Evanston, IL

13 [7] Division of Pulmonary and Critical Care Medicine, Endeavor Health, Evanston, IL

14 [8] Department of Physics and Astronomy, Northwestern University, Evanston, IL

15 [9] Department of Medicine, Northwestern University Feinberg School of Medicine, Chicago, IL

16 [10] NSF-Simons National Institute for Theoretical and Mathematical Biology, Chicago, IL

17

18 <sup>†</sup> Current affiliation: Vizient, Inc. Chicago, IL

19 \* Corresponding Author: Luís A. Nunes Amaral ([amaral@northwestern.edu](mailto:amaral@northwestern.edu))

20

21

22

23

24

25

26

## 27 **Abstract**

28           Physicians could greatly benefit from automated diagnosis and prognosis tools to help address  
29 information overload and decision fatigue. Intensive care physicians stand to benefit greatly from such  
30 tools as they are at particularly high risk for those factors. Acute Respiratory Distress Syndrome (ARDS)  
31 is a life-threatening condition affecting >10% of critical care patients and has a mortality rate over 40%.  
32 However, recognition rates for ARDS have been shown to be low (30-70%) in clinical settings. In this  
33 work, we present a reproducible computational pipeline that automatically adjudicates ARDS on  
34 retrospective datasets of mechanically ventilated adult patients. This pipeline automates the steps outlined  
35 by the Berlin Definition through implementation of natural language processing tools and classification  
36 algorithms. We train an XGBoost model on chest imaging reports to detect bilateral infiltrates, and  
37 another on a subset of attending physician notes labeled for the most common ARDS risk factor in our  
38 data. Both models achieve high performance—a minimum area under the receiver operating characteristic  
39 curve (AUROC) of 0.86 for adjudicating chest imaging reports in out-of-bag test sets, and an out-of-bag  
40 AUROC of 0.85 for detecting a diagnosis of pneumonia. We validate the entire pipeline on a cohort of  
41 MIMIC-III encounters and find a sensitivity of 93.5% — an extraordinary improvement over the 22.6%  
42 ARDS recognition rate reported for these encounters — along with a specificity of 73.9%. We conclude  
43 that our reproducible, automated diagnostic pipeline exhibits promising accuracy, generalizability, and  
44 probability calibration, thus providing a valuable resource for physicians aiming to enhance ARDS  
45 diagnosis and treatment strategies. We surmise that proper implementation of the pipeline has the  
46 potential to aid clinical practice by facilitating the recognition of ARDS cases at scale.

## 47 **Introduction**

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

56           In this study, we report the development and validation of a ML pipeline to help physicians  
57 adjudicate acute respiratory distress syndrome (ARDS). ARDS, a syndrome of severe acute hypoxemia  
58 resulting from inflammatory lung injury<sup>5,6</sup>, is an ideal case for the development of a diagnostic aid tool.  
59 ARDS recognition requires physicians to synthesize information from multiple distinct data streams and  
60 determine whether it fits a standard definition of ARDS. The criteria of the clinically-based Berlin  
61 Definition of ARDS include quantitative data ( $p_aO_2:F_iO_2 \leq 300$  mm Hg), unstructured data (bilateral  
62 opacities on chest imaging), and subjective data (assessing for the presence of ARDS risk factors and  
63 cardiac failure)<sup>5</sup>. Despite ARDS' high prevalence, morbidity, and mortality, prior research has shown that  
64 many patients with ARDS are not recognized by their treating physicians<sup>6,7</sup>. The poor recognition rate of  
65 ARDS is at least partially due to the difficulty in evaluating the Berlin Definition criteria, which requires  
66 the physician to access laboratory data, chest images or radiology notes, other physicians' notes, and  
67 echocardiographic data or notes, then apply the Berlin criteria to determine whether ARDS is present.  
68 Under-recognition of ARDS plays an important role in under-utilization of evidence-based ARDS  
69 treatment (e.g., low tidal volume ventilation and prone positioning), even when physicians believe these  
70 interventions are warranted<sup>8</sup>. An automated approach to help identify ARDS diagnostic criteria has the  
71 potential to be a powerful aid to physician decision-making, leading to improved ARDS recognition and  
72 therefore improved ARDS management.

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

86

## 87 **Data and Methods**

### 88 **Cohort Development and Data Collection**

89 We developed three patient cohorts for this study: Cohort MC1-T1, Cohort MC1-T2, and Cohort  
90 MC2-T3, where MC indicates the “Medical Center” and T indicates the “Time Period.” In addition, we  
91 obtained data from the Medical Information Mart for Intensive Care (MIMIC-III) database<sup>16,17</sup>. MIMIC-  
92 III is a large, single-center database of critically ill patients at a tertiary care medical center. It includes all  
93 the components necessary to identify ARDS and therefore apply our pipeline, and the data is freely  
94 available. For all four cohorts, patients were included if they were at least 18 years old; were admitted to  
95 an adult ICU; and had acute hypoxemic respiratory failure requiring invasive mechanical ventilation (at  
96 least one recorded  $P_aO_2/F_iO_2 \leq 300$  mm Hg while receiving positive-end expiratory pressure  $\geq 5$  cm H<sub>2</sub>O)<sup>7</sup>.  
97 The study was approved by the Institutional Review Boards of Northwestern University (STU00208049)  
98 and Endeavor Health (EH17-325). Table 1 summarizes the data collected for the four cohorts.

99

100 **Table 1.** Cohort characteristics and data annotation. MC refers to “Medical Center”, T refers to “Time  
101 Period”.

Source	Report type	Number	% bilateral infiltrates	Rater IDs
MC1-T1	Chest imaging reports	5,839 reports (800 patients)	60%	1
MC1-T1	Attending physician notes	12,582 notes (790 patients)		
MC1-T1	Adjudicated Attending physician notes	2,034 notes (400 patients)		1, 5
MC1-T1	Echocardiogram reports	1,006 reports (681 patients)		
MC1-T2	Chest imaging reports	6,040 reports (749 patients)	44%	1
MC2-T3	Chest imaging reports	631 reports (90 patients)	34%	1, 2, 3, 4
MIMIC-III	Chest imaging reports	975 reports (100 patients)	22%	1,6
MIMIC-III	Attending physician notes	887 notes (100 patients)		1,6
MIMIC-III	Echocardiogram reports	89 reports (100 patients)		1,6

102

103 Cohort MC1-T1

104 We previously characterized a cohort of 943 patients, which we denote here as MC1-T1, who met  
105 the above inclusion criteria at a single academic medical center between June and December 2013. We  
106 collected the following data: all  $P_aO_2/F_iO_2$  ratios; the unstructured text of all radiologist reports for chest  
107 imaging (radiographs and CT scans), critical care attending physician notes, and echocardiogram reports;  
108 and B-type natriuretic peptide (BNP) values obtained from hospital admission to the earliest of  
109 extubation, death, or discharge. Data were reviewed by study personnel to determine whether each

110 individual Berlin Definition criterion was present, and whether all criteria taken together were consistent  
111 with a diagnosis of ARDS <sup>7</sup>.

112 We collected 5,839 chest imaging reports from 800 Cohort MC1-T1 patients. Study personnel  
113 adjudicated 57% of these chest imaging reports as describing bilateral infiltrates consistent with the Berlin  
114 Definition<sup>5</sup>. We developed our machine learning (ML) approach to bilateral infiltrate adjudication using  
115 these Cohort MC1-T1 reports.

116 For 790 of the 800 Cohort MC1-T1 patients with a chest imaging report, we also had at least one  
117 attending physician note. We collected 12,582 attending physician notes for these patients, of which 2,034  
118 notes from a subset of 400 patients were annotated by study personnel for the presence of ARDS risk  
119 factors (e.g., pneumonia, sepsis, aspiration, etc.)<sup>5</sup>. We used this annotated subset of 2,034 notes to develop  
120 our ML and regular expression (regex) approach for finding ARDS risk factors and cardiac failure  
121 language in attending physician notes.

122 We collected 1,006 echocardiogram (echo) reports from 681 Cohort MC1-T1 patients. Study  
123 personnel from a prior analysis<sup>7</sup> text-matched and adjudicated each echo report for the presence or  
124 absence of: left ventricular ejection fraction < 40%, cardiopulmonary bypass at time of echo, left  
125 ventricular hypertrophy, left atrial dimension > 4cm or left atrial volume index > 28 mL/m<sup>2</sup>, and Grade II  
126 or III diastolic dysfunction. Separately, 35 BNP values were included for 32 patients in Cohort MC1-T1.  
127 We used echo reports and BNP values to develop our objective cardiac failure rule-out approach.<sup>7</sup>

128 Since identification of Berlin Definition-consistent bilateral infiltrates is the most challenging  
129 task<sup>11,18</sup> in our computational pipeline, we analyzed the chest imaging reports from two additional cohorts  
130 of patients to test our ML algorithm, Cohort MC1-T2 and Cohort MC2-T3.

### 131 Cohort MC1-T2

132 Cohort MC1-T2 comprises 749 patients admitted during 2016 at the same medical center as  
133 Cohort MC1-T1 and meeting the same inclusion criteria. We collected 6,040 chest imaging reports for  
134 these patients. Study personnel adjudicated 44% of Cohort MC1-T2 reports as describing bilateral  
135 infiltrates. We used this cohort only to train a second ML algorithm for bilateral infiltrate adjudication.

136 Cohort MC2-T3

137 Cohort MC2-T3 comprises 90 patients admitted to a different medical center in 2017–2018 and  
138 meeting the same inclusion criteria as Cohorts MC1-T1 and MC1-T2. We collected 631 chest imaging  
139 reports for these 90 patients. Study personnel adjudicated 34% of these chest imaging reports as  
140 describing bilateral infiltrates. We used these reports from Cohort MC2-T3 only to test ML algorithms for  
141 bilateral infiltrate adjudication.

142 MIMIC-III

143 We identified the set of patients in the MIMIC-III dataset who satisfied the inclusion criteria used  
144 to develop cohort MC1-T1. This resulted in a set comprising 3,712 encounters. We then used our  
145 automated pipeline to adjudicate the presence or absence of ARDS for all those encounters, and randomly  
146 selected a balanced cohort comprising 100 encounters, which we denote as the MIMIC-III cohort. Each of  
147 the encounters in the MIMIC-III cohort was adjudicated by one critical care physician and one internal  
148 medicine physician for whether each individual Berlin Definition criterion was present, and whether all  
149 criteria taken together were consistent with a diagnosis of ARDS. This cohort, with physician  
150 adjudications, is publicly available at Northwestern’s ARCH database.

151 The records of these 100 MIMIC-III patients included 975 chest imaging reports, 887 attending  
152 physician notes, and 89 echocardiogram (echo) reports. The critical care physician adjudicated 22.3% of  
153 these chest imaging reports as describing bilateral infiltrates consistent with the Berlin Definition<sup>5</sup>. The  
154 same individual also annotated 887 attending physician notes for the presence of ARDS risk factors<sup>5</sup> and  
155 cardiac failure language, and 89 echo reports for the presence or absence of: left ventricular ejection  
156 fraction < 40%, cardiopulmonary bypass at time of echo, left ventricular hypertrophy, left atrial  
157 dimension > 4cm or left atrial volume index > 28 mL/m<sup>2</sup>, and Grade II or III diastolic dysfunction.<sup>7</sup> We  
158 used these adjudicated datasets to evaluate the performance of our ML algorithm.

159 **Analysis**

160 Adjudication of bilateral infiltrates from chest imaging reports

161 We preprocessed chest imaging reports to remove patient information, non-informative sections  
162 (e.g. technique, indication, history, etc.), and non-informative words. We then tokenized the remaining  
163 sections (i.e., separated the text into sets of unigrams and bigrams) and prepared the data for use of a “bag  
164 of words” approach (i.e, we vectorized these tokens according to their counts in the imaging reports).  
165 When training a ML model on a given corpus, we used the 200 most frequently appearing tokens across  
166 the imaging reports in the respective corpus as model features.

167 We trained four different ML models (decision trees, logistic regression, random forest  
168 classifiers, and extreme gradient boosting ‘XGBoost’<sup>19</sup>) on chest imaging reports from Cohort MC1-T1,  
169 which was also used to perform hyperparameter tuning for the four models. We performed  
170 hyperparameter tuning using Bayesian optimization, which is available through the *hyperopt* package  
171 (v.0.2.7)<sup>20</sup> for Python (v.3.10.12). For each model, we performed 5-fold cross-validation to obtain the  
172 mean Area under the Receiver Operating Characteristic (AUROC) curve for each hyperparameter  
173 combination considered. We then selected the optimal combination of hyperparameters as the one  
174 yielding the highest 5-fold cross-validation mean AUROC after at least 100 iterations. We also derived  
175 another set of optimal hyperparameters for XGBoost trained with chest imaging reports from Cohort  
176 MC1-T2 in the same fashion.

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

183 For comparing the performance of the four models, we used nested cross-validation by doing 5-  
184 fold cross-validation to obtain a mean AUROC across five different folds. Furthermore, each fold’s  
185 training set was used to tune that model’s hyperparameters as described above (i.e., five separate  
186 hyperparameter tuning exercises). However, we note that we used a 3-fold cross-validation strategy for



187 this tuning due to the computational cost of running nested cross-validation. We repeated this nested cross  
188 validation 10 times, each time with a different resampling with replacement of the data (i.e. a bootstrap)  
189 to yield a distribution of mean AUROCs on test sets.

190 We used 95% confidence intervals to compare ROC curves and AUROCs across different  
191 models. For obtaining feature/token importance during training, we employed the default “importance”  
192 method that version 1.1.3 of the scikit-learn package implements for decision tree, logistic regression, and  
193 random forest algorithms, and version 1.7.4 of *xgboost* package for XGBoost. For decision tree and  
194 random forest, feature importance corresponds to the mean decrease in Gini impurity; for logistic  
195 regression, importances correspond to the mean value of coefficients in the fitted linear equation; and for  
196 XGBoost, the importance corresponds to the mean gain in predictive performance obtained by including a  
197 particular feature in the trees.

198 We also evaluated the inter-rater disagreement rate for chest imaging reports from MIMIC III.  
199 For this purpose, we obtained two independent adjudications (one critical care physician and one internal  
200 medicine physician) for 975 reports available, and split imaging reports into three groups according to  
201 XGBoost output probabilities. For each group, we then calculated the fraction of imaging reports for  
202 which these independent raters disagreed on their adjudications.

203 Finally, to assess how our XGBoost implementation for chest imaging reports generalizes to other  
204 datasets, we tuned hyperparameters and then trained XGBoost models on all chest imaging reports from  
205 Cohort MC1-T1 and MC1-T2, separately. We then tested each of the two models on the two other chest  
206 imaging corpora the model had not yet seen by comparing the AUROC values. We used 100 bootstrapped  
207 samples to gather 95% confidence intervals for the mean AUROC values.

#### 208 Adjudication of risk factors from physician notes

209 The Berlin Definition of ARDS requires the presence of at least one risk factor — e.g.,  
210 pneumonia, sepsis, shock, inhalation, pulmonary contusion, vasculitis, drowning, drug overdose — within  
211 seven days of non-cardiogenic acute respiratory failure. We preprocessed attending physician notes from  
212 Cohort MC1-T1 to remove identifiable information from the text of these notes. We then used regular

213 expressions (regex v2022.10.31) to match keywords related to risk factor and heart failure language (see  
214 SI: Regular expression list 1, for a complete list of risk factors). To validate this strategy, we ensured that  
215 this regex approach matched 100% of the notes that had a positive adjudication for a particular risk factor  
216 (or close to 100% as possible). We also corrected common spelling errors on important keywords, such as  
217 ‘pneunonia’, ‘spetic’ or ‘cardigenic’. To prevent data leakage during ML development, we again split the  
218 adjudicated notes into train and test sets by encounter, not note.

219         Adjudicating the presence of a risk factor is not as simple as finding a particular keyword in a  
220 physician note. For example, a note stating “patient is unlikely to have pneumonia” should not be  
221 classified as evidence of pneumonia. To account for such possibilities, we implemented a strategy in  
222 which, after matching a particular keyword, we extract a text string from the note starting 100 characters  
223 prior to the occurrence of the keyword and extending 100 characters post the keyword. Subsequently, we  
224 tokenized and vectorized the strings as described in the previous subsection.

225         Using the vectorized tokens, we trained XGBoost models for a select group of risk/cardiac failure  
226 factors, employing a similar nested cross-validation strategy as the one pursued for the adjudication of  
227 chest imaging reports (except in this case we used 100 resamples instead of 10). Note that not all risk  
228 factors were amenable to a ML approach: we chose risk factors that had more than 100 notes annotated,  
229 were risk factors for ARDS (or a cardiac failure criterion) and had relatively balanced yes/no proportions  
230 after regex-matching (between 33% and 66%, see SI Table 1). This resulted in the use of 1409  
231 adjudicated attending notes from 337 patients for ML development (see SI Table 1 for a breakdown of  
232 notes used for each risk factor/cardiac failure criterion).

### 233 Objective adjudication of cardiac failure from echocardiogram reports

234         We preprocessed echo reports from MC1-T1 to remove identifiable information from text. We  
235 then developed regex patterns that first matched keywords associated with the following parameters: left  
236 ventricular ejection fraction, cardiopulmonary bypass, left atrial diameter, left atrial volume index, left  
237 ventricular hypertrophy, and grade II or III diastolic dysfunction (see SI: Regular expression list 2 for  
238 regex patterns). Once these parameters were found in the echo report text, we then extracted numerical

239 values of numerical variables (left ventricular ejection fraction, left atrial diameter, and left atrial volume  
240 index), and the matched text otherwise.

#### 241 Design of ARDS adjudication pipeline: chaining Berlin Definition steps

242 Each of the steps outlined above automates the adjudication of specific criteria in the Berlin  
243 Definition, with the modifications specified previously<sup>7</sup>. We integrate the different criteria into a single  
244 pipeline to build an automated ARDS adjudication pipeline.

245 The ARDS adjudication pipeline first flags encounters with at least one hypoxemia measurement  
246 (i.e., one instance of PF ratio  $\leq 300$  mm Hg while PEEP  $\geq 5$  cm H<sub>2</sub>O), and then uses the predictions of an  
247 XGBoost model trained on chest imaging reports from MC1-T1 to adjudicate presence of bilateral  
248 infiltrate language in chest imaging reports. Upon settling these two criteria, the pipeline flags whether  
249 the hypoxemia record and the report consistent with bilateral infiltrates have timestamps within 48 hours  
250 of each other (which we term “qualified hypoxemia”). In addition, at this step the pipeline ensures that the  
251 hypoxemia record was taken at or after intubation. Next, the pipeline uses the predictions of an XGBoost  
252 model trained on attending physician notes that have pneumonia keywords to adjudicate pneumonia on all  
253 notes and uses regex to flag presence of other risk factors, cardiac failure language (e.g., cardiac arrest),  
254 and indicators of cardiogenic and noncardiogenic language. Finally, the pipeline flags whether an  
255 attending physician note has a timestamp that falls between one day prior to and seven days after the latter  
256 timestamp of any of the qualifying hypoxemia-bilateral infiltrates pairs. Once these annotations are  
257 integrated, the pipeline proceeds to adjudicate whether an ARDS diagnosis is warranted.

258 If any risk factor is identified in this time window for an encounter, via XGBoost or regex, the  
259 pipeline adjudicates the encounter as an ARDS case. If no risk factors are identified, but cardiac failure  
260 language is identified in the notes through the use of regex, the pipeline adjudicates the encounter as a  
261 “No ARDS” case.

262 For all other encounters not meeting the risk factor or cardiac failure language criteria described  
263 above, the pipeline flags the case for objective cardiac failure assessment<sup>7</sup>. This assessment is done  
264 sequentially instead of by flagging. If any encounter had BNP greater than 100 pg/mL (an indicator of

265 heart failure), the pipeline adjudicates “No ARDS” for such an encounter. The pipeline then considers the  
266 remaining encounters for each subsequent criteria, adjudicating “No ARDS” if the encounter had any of  
267 the following: left ventricular ejection fraction < 40%, cardiopulmonary bypass found in echocardiogram  
268 report, or at least two of the following present in the echocardiogram report: (i) left atrial diameter > 4 cm  
269 or left atrial volume index > 28 mL/m<sup>2</sup>, (ii) left ventricular hypertrophy, or (iii) Grade II or III diastolic  
270 dysfunction.

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

274

## 275 **Results**

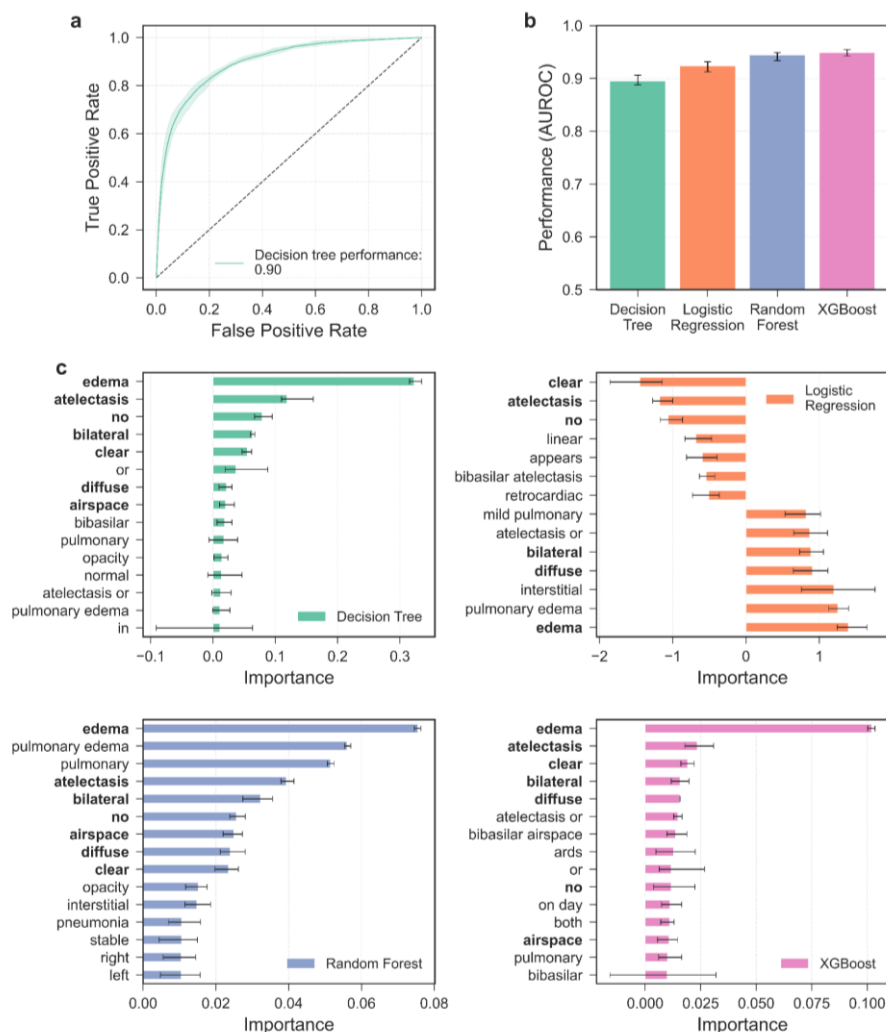
### 276 **Adjudication of bilateral infiltrates**

277 Figure 1a shows the Receiver Operator Characteristic (ROC) curves for the decision tree model  
278 applied to chest imaging reports from Cohort MC1-T1. We quantify model predictive performance using  
279 the areas under the ROC curves (AUROCs). We observe that once hyperparameters for each model are  
280 optimized, all models trained on chest imaging reports from MC1-T1 achieve AUROCs of at least 0.90  
281 on the training set (Decision tree: AUROC = 0.89, 95% CI = [0.89, 0.91]; logistic regression: AUROC =  
282 0.92, 95% CI = [0.91, 0.93]; random forest: AUROC = 0.94, 95% CI = [0.93, 0.95]; XGBoost: AUROC =  
283 0.95, 95% CI = [0.94, 0.95]) (Fig. 1b).

284 We next calculated the importance that each model assigned to the 200 tokens used as features.  
285 Reassuringly, we find that the four models consistently identify tokens such as edema, bilateral, clear, and  
286 atelectasis as the most predictive (Fig. 1c). These tokens correspond closely to the inclusion/exclusion  
287 language we developed to address Berlin Definition shortcomings<sup>7</sup>, which we also observed when  
288 implementing Shapley-additive explanations (SHAP) values to assess feature importance (SI Fig. 1).

289

290 **Figure 1. Machine learning (ML) models achieve high-performance in adjudicating the presence of bilateral**  
 291 **infiltrates from chest imaging reports.** Error bars show 95% confidence intervals for estimates of the mean  
 292 obtained using bootstrapping. **a)** Receiver operating characteristic (ROC) curve for the decision tree model trained  
 293 on chest imaging reports from Cohort MC1-T1. **b)** Bootstrapped mean area under the ROC (AUROC) show that all  
 294 four ML approaches yield AUROCs greater or equal to 0.90. **c)** Feature importances for the four different ML  
 295 approaches considered. Features in bold are highly ranked in importance in at least 3 of the 4 approaches.

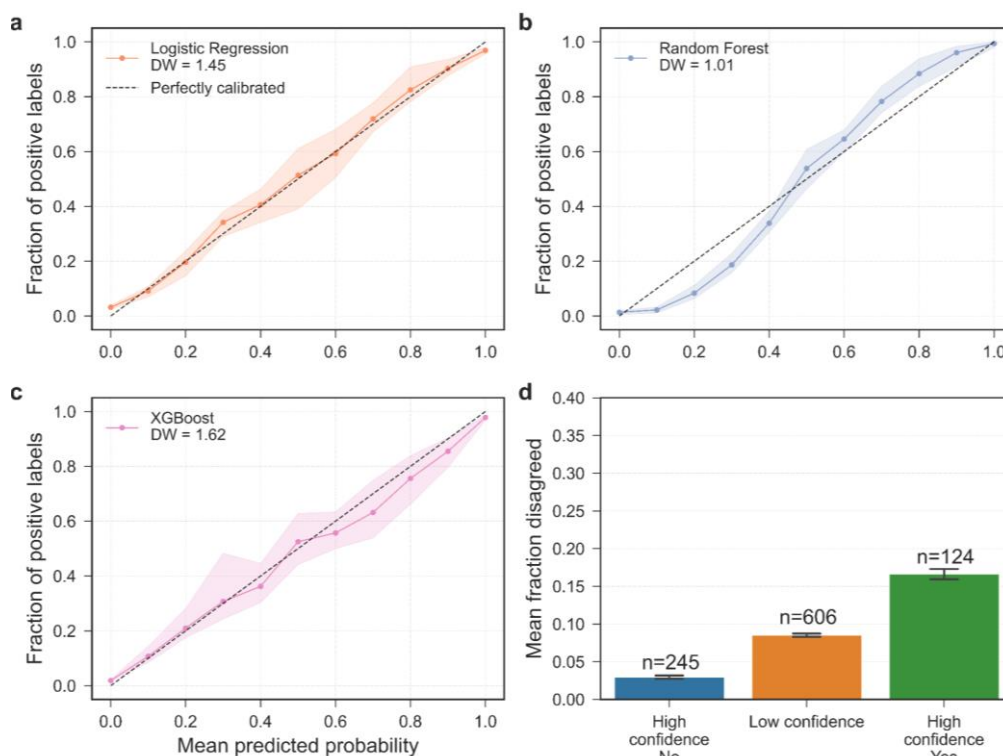


296  
 297  
 298 We then assessed how calibrated were the outcome probabilities from the models by comparing  
 299 model outcome probability after training to actual probability of occurrence in considered encounters  
 300 from MC1-T1 chest imaging reports. Figures 2a-c suggest that logistic regression and XGBoost models  
 301 output probabilities that are well calibrated, which is expected given their use of similar loss functions for

302 fitting (log-loss). In contrast, random forest produces poorly calibrated probabilities, being over-confident  
303 when forecasting with confidence levels lower than 50%, and under-confident with confidence levels  
304 greater than 50%. We thus select XGBoost as the implemented approach for our pipeline as it offers the  
305 highest predictive performance (AUROC = 0.95, 95% CI = [0.94, 0.95]) and well-calibrated forecasts on  
306 the training set (Durbin-Watson statistic = 1.55, 95% CI = [1.10-1.87]).

307

308 **Figure 2. Assessment of ML implementation probabilities.** Comparing calibration of MC1-T1 probabilities by a)  
309 logistic regression, b) random forest, and c) XGBoost. A perfectly calibrated model would have a 1:1 relationship  
310 between fraction of positive labels and mean probabilities (i.e., it would overlay the diagonal line). The Durbin-  
311 Watson statistic, DW, probes for correlations in the residuals, if DW is close to 2, then one can rule out correlations  
312 in the residuals, implying good linear behavior. d) Comparing inter-rater disagreement rate to the confidence in  
313 adjudicating bilateral infiltrates from chest imaging reports from MIMIC III by an XGBoost model trained on chest  
314 imaging reports from Cohort MC1-T2.



315

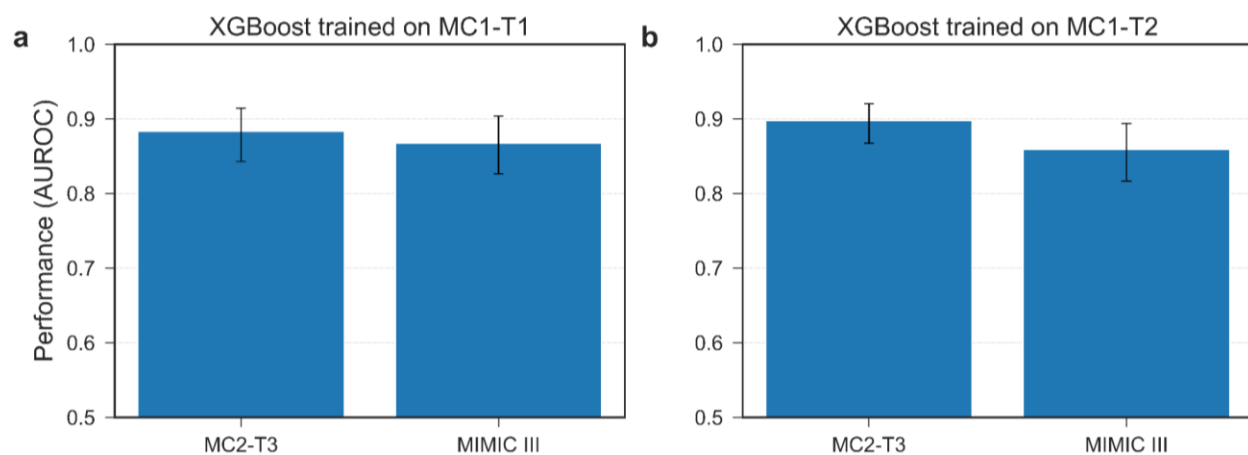
316

317 The good calibration of the prediction of the XGBoost model prompted us to test the hypothesis

318 that the estimated probabilities by the model may contain information regarding the agreement rate of

319 physician adjudications of the same chest imaging reports. To this end, we leveraged MIMIC III report  
320 annotations by two independent raters, a critical care physician and an internal medicine physician. We  
321 evaluated how the agreement between these two independent raters in adjudicating chest imaging reports  
322 from MIMIC III associated with the output probabilities of an XGBoost classifier trained on chest  
323 imaging reports from Cohort MC1-T1. Specifically, we binned predictions into three groups: high-  
324 confidence ‘No’ (probability of bilateral infiltrates <10%), high-confidence ‘Yes’ (probability of bilateral  
325 infiltrates >90%), low confidence (all other probabilities of bilateral infiltrates). As seen on Figure 2d, the  
326 interrater disagreement was highest (16.6%) for the cases of high confidence ‘Yes’ predictions. This  
327 suggests that the model could be an effective way to avoid false negatives by a physician’s  
328 misinterpretation of chest imaging reports and can be used to alert that there is a high chance of a report  
329 being consistent with bilateral infiltrates. It also suggests that this model is most reliable when it indicates  
330 a low probability of a chest imaging report showing bilateral infiltrates.

331  
332 **Figure 3. Evaluation of XGBoost’s generalization performance to MC2-T3 and MIMIC-III cohorts.** We show  
333 AUROCs with bootstrapped 95% confidence intervals as error bars for XGBoost models trained on a) chest imaging  
334 reports from cohort MC1-T1, and b) chest imaging reports from cohort MC1-T2.

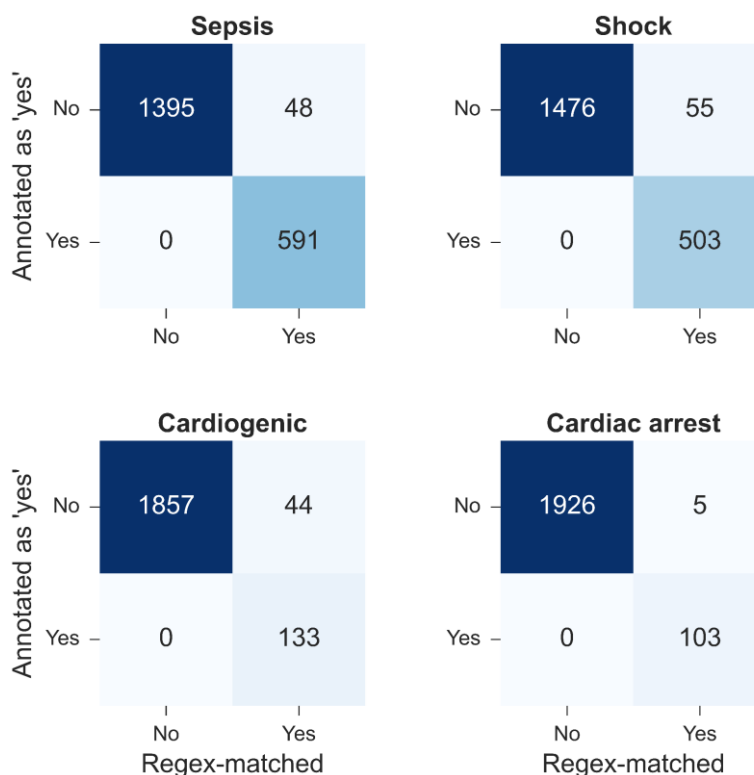


336  
337  
338 To assess how a model developed for a specific cohort generalizes to a different health system  
339 dataset, we tested this MC1-T1-trained XGBoost model on chest imaging reports from Cohort MC2-T3

340 and MIMIC-III (Fig. 3a). The MC1-T1-trained XGBoost model yields AUROCs of 0.88 (95%CI = [0.84,  
341 0.91]) and 0.87 (95%CI = [0.83, 0.90]) when applied to Cohort MC2-T3 and MIMIC-III, respectively.  
342 We also trained a second XGBoost model on chest imaging reports from Cohort MC1-T2 and tested this  
343 model against chest imaging reports from MC2-T3 and MIMIC-III. We found similar results to the first  
344 XGBoost model, with an AUROC of 0.90 (95%CI = [0.87, 0.92]) when applying this MC1-T2-trained  
345 model to chest imaging reports from MC2-T3, and an AUROC of 0.86 (95%CI = [0.82, 0.89]) when  
346 applying this model to chest imaging reports from MIMIC-III (Fig. 3b).

347

348 **Figure 4. Confusion matrices for the performance of regex approach to capture risk factors in attending**  
349 **physician notes.** ‘Sepsis’ and ‘shock’ are the most prevalent risk factors for ARDS after pneumonia. ‘Cardiogenic’  
350 and ‘cardiac arrest’ are the most prevalent cardiac failure keywords in attending physician notes. Notice the absence  
351 of false negatives, which indicates that regex-matching can capture all instances in which a physician adjudicated  
352 the language as being present in the attending physician note.



353

354

355



## 356 **Extracting ARDS risk factors in attending physician notes**

357           We first developed regex patterns to match keywords for the risk factors. As shown in Fig. 4, the  
358 developed regex patterns match 100% of the notes that were annotated ‘yes’ for a particular risk factor in  
359 MC1-T1. When we count the total number of notes annotated as either yes or no, the most prevalent  
360 matches were sepsis (744 notes annotated vs. 748 notes regex-matched), pneumonia (636 notes annotated  
361 vs. 955 regex-matched), and shock (604 notes annotated vs. 607 regex-matched). For the cardiac failure  
362 criteria, the relevant matches were the cardiogenic keyword (176 notes annotated vs. 725 regex-matched)  
363 to qualify the matching of shock, and congestive heart failure (254 notes annotated vs. 352 regex-  
364 matched). We thus feel confident that the built regex-patterns can match nearly the entirety of the notes  
365 annotated as “yes” for specific risk factors.

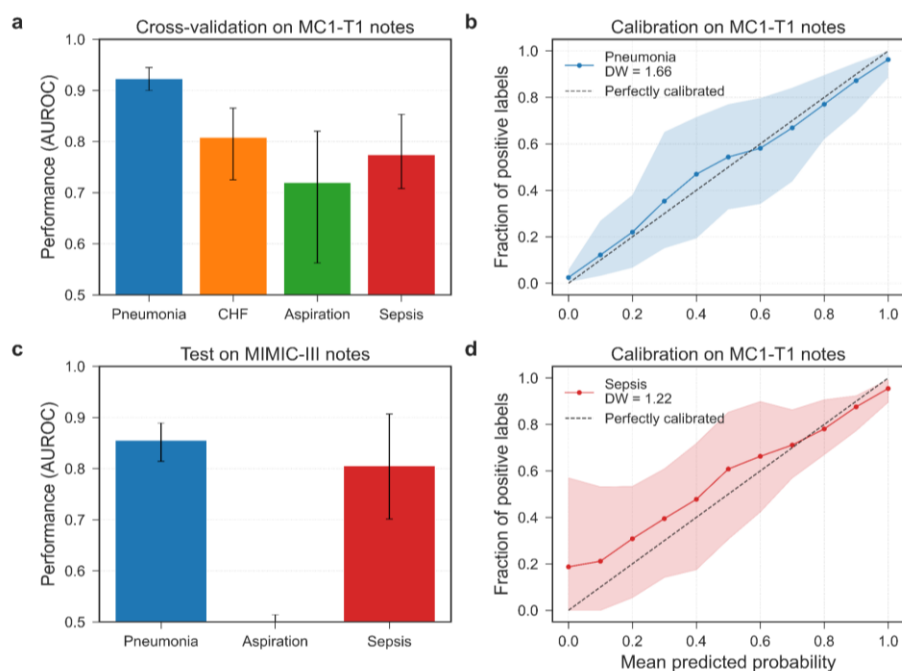
366           Next, we trained an XGBoost model on attending physician notes from Cohort MC1-T1 to  
367 adjudicate pneumonia, aspiration, congestive heart failure, and sepsis. We chose these risk factors for ML  
368 because at least 100 attending physician notes were annotated for them, and their annotations have  
369 relatively balanced yes/no proportions after regex-matching (between 33% and 66%, with the exception  
370 of sepsis; see Table S1). We did not use all 2,034 records from Cohort MC1-T1 to train each of the  
371 models since not every record had an annotation or a keyword for a given risk factor. For instance, only  
372 636 notes in Cohort MC1-T1 included an annotation for pneumonia, whereas we were able to match 955  
373 notes for pneumonia using regex. Therefore, our training dataset for each of the three models consisted of  
374 all notes that were regex-matched for that particular risk factor. For notes that were regex-matched but did  
375 not have an annotation, we imputed the annotation as ‘No’, or zero.

376           We used nested cross-validation for every XGBoost implementation (pneumonia, sepsis, etc.).  
377 This involved splitting data into a train and test set, tuning hyperparameters on the train set using 3-fold  
378 cross validation and measuring AUROC on the test set. Since we used 5-fold cross-validation, this  
379 process was repeated 5 times per bootstrapped sample, yielding 5 AUROCs. We repeated the above  
380 process for a total of 100 bootstrapped data samples to evaluate the mean AUROCs obtained by each of  
381 the models. We observed that out of the three XGBoost implementations, the pneumonia model yielded

382 the best discriminative performance during cross-validation (Pneumonia: AUROC = 0.93, 95% CI = [0.90,  
383 0.95]; CHF: AUROC = 0.82, 95% CI = [0.75, 0.88]; Aspiration: AUROC = 0.73, 95% CI = [0.60, 0.83];  
384 Sepsis: AUROC = 0.76, 95% CI = [0.66, 0.85]; Fig. 5a), the best calibration (DW = 1.70, 95% CI = [1.13-  
385 2.22]; Fig. 5b), and the overall better generalizability to attending notes from MIMIC-III (Pneumonia:  
386 AUROC = 0.86, 95% CI = [0.81, 0.89]; Aspiration: AUROC = 0.42, 95% CI = [0.31, 0.51]; Sepsis:  
387 AUROC = 0.81, 95% CI = [0.70, 0.91]; Fig. 5c). This is in stark contrast to the XGBoost models for  
388 congestive heart failure, aspiration, and sepsis: These models have underwhelming performance on cross-  
389 validation, generalizability to MIMIC-III, and turn out poorly calibrated on attending notes from MC1-T1  
390 (Fig. 5d). Thus, we decided against integrating these three ML models into our pipeline.

391

392 **Figure 5. XGBoost model performance in adjudicating for presence of risk factors in attending physician**  
393 **notes amenable to ML techniques. a)** Cross-validated performance of XGBoost models trained to adjudicate  
394 pneumonia, congestive heart failure, aspiration, and sepsis on MC1-T1 attending notes. **b)** Training set calibration  
395 curve for the pneumonia XGBoost model. **c)** Test set performance of XGBoost models trained to adjudicate  
396 pneumonia, aspiration, and sepsis using MC1-T1 attending notes. We did not have labels for CHF available for  
397 MIMIC-III, therefore we did not explore the generalizability of this model. **d)** Training set calibration curve for the  
398 sepsis XGBoost model.



399

400           Instead, we use regex-matching and a simple heuristic to adjudicate other ARDS risk  
401 factor/cardiac failure language in Cohort MC1-T1. Note that this is not a limitation since many other  
402 types of ARDS risk factor/cardiac failure language are more predictable in their adjudication. For  
403 example, of the 105 attending physician notes matching for ‘cardiac arrest’, 103 were annotated as ‘yes’  
404 (Table S1). Thus, our heuristic was to adjudicate a risk factor as ‘present’ if it was annotated as ‘yes’ in  
405 more than 80% of the matched notes. These included shock, cardiac arrest, pulmonary contusion,  
406 vasculitis, drowning, and overdose.

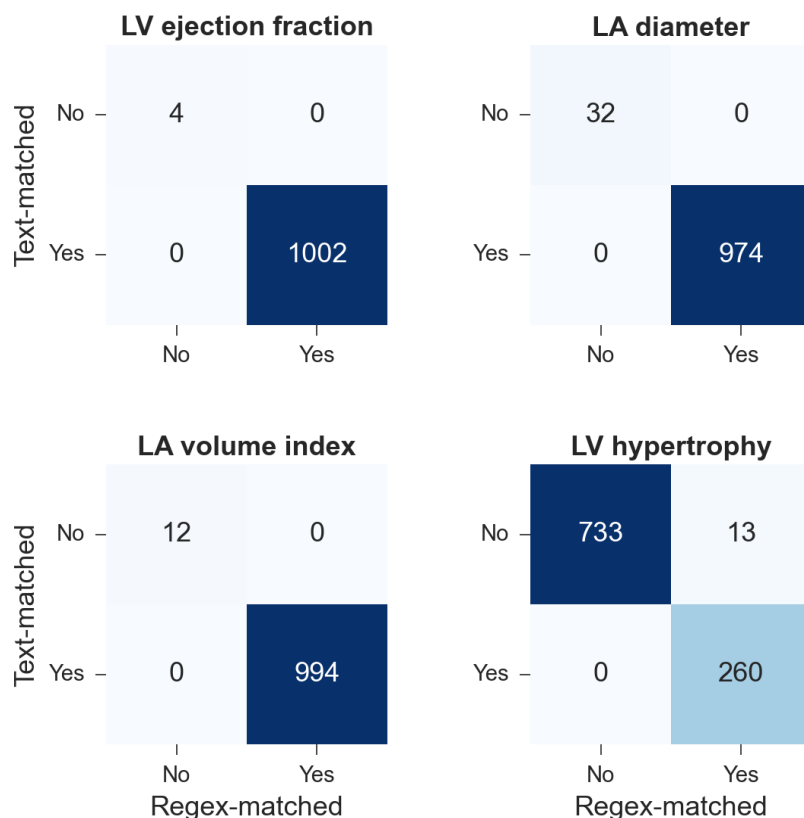
#### 407 **Adjudication of cardiac failure from echocardiogram (echo) reports**

408           The criteria for the objective assessment of cardiac failure rely on the following six factors: left  
409 ventricular ejection fraction, cardiopulmonary bypass, left atrial diameter, left atrial volume index, left  
410 ventricular hypertrophy, and grade II or III diastolic dysfunction. Because echo reports are highly  
411 standardized, it is possible to extract these factors from the reports using regex. Moreover, we had access  
412 to echo reports from Cohort MC1-T1 which were previously text-matched, enabling us to validate our  
413 regex approach.

414           Using the regex patterns listed in the SI, we analyze Cohort MC1-T1’s echo reports for the  
415 presence or absence of each of the six factors of interest. Figure 6 demonstrates that not all six factors  
416 were present in every echo report. For three of the six factors — ‘left ventricular ejection fraction’, ‘left  
417 atrial dimension/diameter’, and ‘left atrial volume index’— we found excellent agreement between regex  
418 and text-matching. Two of the other three, ‘cardiopulmonary bypass’ and ‘diastolic function’, were not  
419 text-matched, so no comparison can be made. For ‘left ventricular hypertrophy’, the regex-matching  
420 procedure correctly captured the desired language, indicating that the original text-matching procedure  
421 failed to identify 13 echo reports. In addition, we validated the numerical values extracted through this  
422 regex approach by randomly selecting 10% of echo reports for visual inspection of values and comparing  
423 against values extracted through regex. We found 100% concordance between values extracted and those  
424 retrieved manually (SI Table 2).

425

426 **Figure 6. Confusion matrices comparing the flagging performance of regex-matching against text-matching**  
427 **for ‘Left ventricular ejection fraction’, ‘Left atrial dimension/diameter’, ‘Left atrial volume index’, and ‘Left**  
428 **ventricular hypertrophy’.** Note the large discrepancy for the annotations of ‘left ventricular hypertrophy’, which is  
429 explained in text.



430

431

432

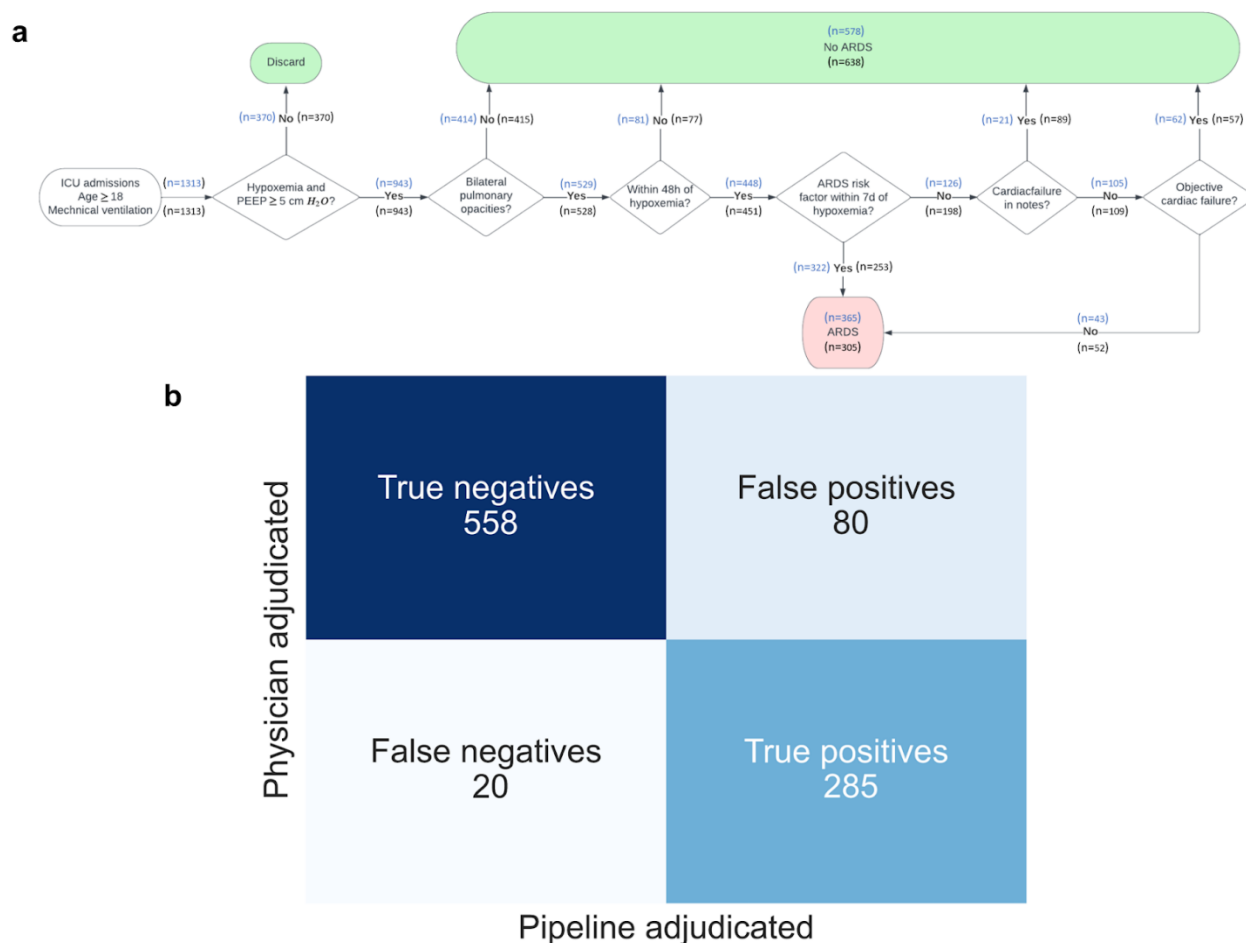
### 433 **Adjudication of ARDS for entire MC1-T1 cohort**

434 We are now ready to compare the performance of our complete pipeline against the previously  
435 reported ARDS adjudication<sup>7</sup>. For the XGBoost model components of the pipeline (chest imaging reports  
436 and pneumonia risk factor adjudications), we use a threshold of 50% to map estimated probabilities into  
437 binary yes/no decisions.

438 We conduct the evaluation of our pipeline for the 943 patients in the MC1-T1 ARDS adjudication  
439 cohort who were 18 years and older, received invasive mechanical ventilation, and had acute hypoxemic  
440 respiratory failure (Fig. 7). 143 patients had no chest imaging report available and were adjudicated as

441 negative for ARDS. The remaining 800 patients had at least one chest imaging report available. The  
 442 XGBoost model trained on MC1-T1 adjudicated bilateral infiltrates within 48 h of a hypoxemic episode  
 443 for 529 patients. Of these 529 patients, 448 had at least one of the qualified hypoxemic events occurring  
 444 post-intubation and 322 had a risk factor within 7 days of the qualified hypoxemia event, and were  
 445 adjudicated as being positive for ARDS.

446  
 447 **Figure 7. Machine learning computational pipeline for adjudication of MC1-T1 cohort yields a small fraction**  
 448 **of false negatives and a manageable fraction of false positives.** a) Flowchart of ARDS diagnosis by  
 449 computational pipeline (blue) vs. physician (black). b) Confusion matrix comparing physician adjudication from  
 450 previous publication<sup>7</sup> against ML computational adjudication pipeline.

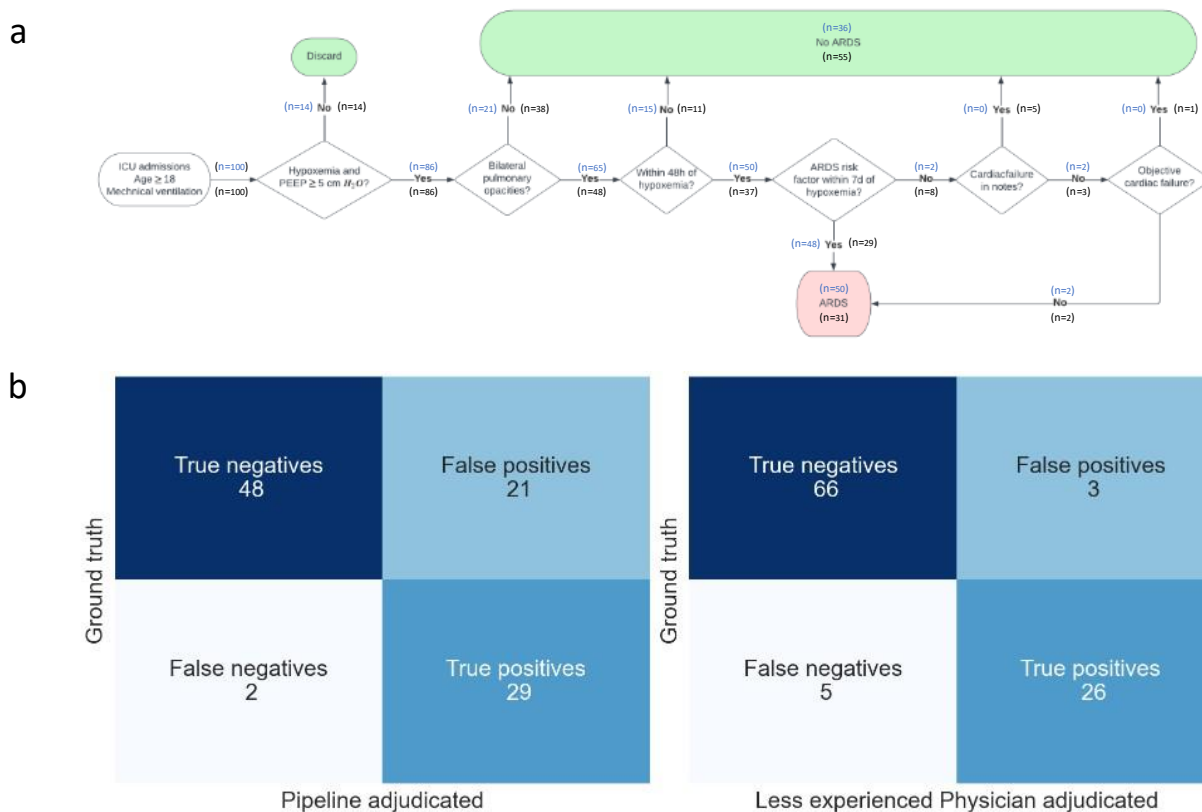


451  
 452  
 453 The remaining 126 patients were then evaluated for cardiac failure. For 21 patients, the physician  
 454 notes indicated cardiac failure and they were adjudicated as negative for ARDS. The last 105 patients

455 were then adjudicated using the objective cardiac failure assessment step; 62 were adjudicated to have  
 456 cardiac failure and thus negative for ARDS and the remaining 43 were adjudicated as positive for ARDS.  
 457 In total, the pipeline adjudicated 365 patients as positive for ARDS and 578 as negative for ARDS.

458 In summary, using a simple 50% probability cutoff for both ML algorithms, our pipeline yields  
 459 close agreement with the physician adjudication of ARDS for this cohort<sup>7</sup> (Fig. 7a). Specifically, the  
 460 pipeline yields a sensitivity or true positive rate of 93.4% on this cohort, which compares favorably to the  
 461 19% ARDS diagnosis rate we found on this cohort<sup>7</sup>. Importantly, this high sensitivity is achieved while  
 462 maintaining a very low 12.5% rate of false positives.

463  
 464 **Figure 8. Machine learning computational pipeline for adjudication of MIMIC-III cohort yields a small**  
 465 **fraction of false negatives and a manageable fraction of false positives. a)** Flowchart of ARDS diagnosis by  
 466 computational pipeline (blue) vs. physician (black). **b)** Confusion matrix comparing physician adjudication (ground  
 467 truth) against ML computational adjudication pipeline (left panel), and physician adjudication (ground truth)  
 468 against a less experienced physician adjudication.



469

## 470 **Adjudication of ARDS for MIMIC-III labeled subset**

471 We applied our automated ARDS adjudication to the 100 patient encounters in the MIMIC-III  
472 cohort. We then compared physician adjudication against the pipeline's (Fig. 8). Reassuringly, and again  
473 using a simple 50% probability cutoff for both ML algorithms, we find that the overall performance of  
474 our pipeline on the MIMIC-III cohort subset is strikingly similar to its performance on the development  
475 cohort<sup>7</sup> (Fig. 8b). Specifically, the pipeline yields a sensitivity or true positive rate of 93.5% on this  
476 cohort, which compares favorably to the 22.6% ARDS documentation rate we found in this subset. This  
477 high sensitivity is achieved while maintaining a relatively low 26.1% rate of false positives. Moreover,  
478 the false negative rate of pipeline adjudication is lower than that of a less experienced physician,  
479 highlighting the pipeline's potential to aid physicians in ARDS diagnosis.

480

## 481 **Discussion**

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

491 In this study, we construct and validate a ML pipeline for automating the adjudication of ARDS  
492 according to the Berlin Definition based on data from EHRs. We constructed high performing decision  
493 tree-based models (XGBoost) to adjudicate chest imaging reports for bilateral infiltrate language and  
494 attending physician notes for the presence of pneumonia. These tree-based methods estimate probabilities,

495 the first stage of any classification problem, and enable physicians to optimize the false positive vs. false  
496 negative tradeoff by adjusting the decision cutoff. In addition, our implementation of XGBoost allows  
497 language-level interpretation of estimated probabilities, which can enhance physician trust in ML models.

498 Supplementing the above XGBoost models with regular expressions to identify other ARDS risk  
499 factors and cardiac failure, and additional structured data to objectively rule out cardiac failure,<sup>7</sup> enabled  
500 the automated adjudication of the complete set of Berlin Definition criteria. This pipeline demonstrated  
501 excellent test characteristics, including false negative and false positive rates of 6.9% and 12.4%,  
502 respectively, at the 50% decision cutoff. We then validated the generalizability of the pipeline on a subset  
503 of the MIMIC-III dataset, demonstrating a similar – high – level of performance.

504 To our knowledge, this study is the first to automate the entire Berlin Definition process using  
505 ML and rules-based methods in a multi-center, open-source, generalizable manner. Previous attempts at  
506 automated ARDS adjudication used single-center EHR data to adjudicate individual Berlin criteria or  
507 used non-reproducible methods. Afshar et. al. used text features in chest imaging reports for ARDS  
508 identification, achieving a maximum AUROC of 0.80 for that task<sup>22</sup>. However, our work identifies ARDS  
509 by considering data beyond chest imaging reports. Sathe et. al. developed EHR-Berlin, evaluating the  
510 Berlin Definition using ML and rules-based methods, but their focus was limited to COVID-19 patients<sup>15</sup>;  
511 by using a cohort of patients who were already defined as having an ARDS risk factor, they effectively  
512 eschewed the need to identify ARDS risk factors or cardiac failure. In contrast, our study considers any  
513 adult patient placed on mechanical ventilation, which requires evaluating all Berlin Definition  
514 components. Finally, Song and Li developed a fully rules-based tool that automates the entire Berlin  
515 Definition, both achieving identical high performance<sup>13,14</sup>. However, their models were constructed within  
516 a single hospital, which may not be reproducible across different health systems. Conversely, the  
517 generalizability and reproducibility of our ARDS adjudication pipeline have been demonstrated  
518 conclusively.

519 Machine learning offers a powerful solution for efficiently analyzing large volumes of data that  
520 would otherwise require countless hours of human effort. A recent study that combined NLP techniques



521 with manual chart abstraction to evaluate clinical trial outcomes confirms this assertion. The time spent  
522 on review decreased from 2,000 laborious hours for manual chart abstraction to just 34.3 hours with NLP-  
523 screened manual chart abstraction<sup>23</sup>. Similarly, our pipeline demonstrates the ability to adjudicate ARDS  
524 for multiple cohorts of hundreds of patients in under five minutes by training XGBoost models at runtime  
525 (with even faster runtimes possible using pre-trained models on inference mode).

526 A potential limitation is that previous studies have shown significant variability in the diagnosis  
527 of ARDS among critical care physicians, particularly in relation to interpretation of chest imaging<sup>18</sup>. We  
528 attempted to mitigate concerns raised by this challenge by only relying on chest imaging reports written  
529 by radiologists. This allowed us to use previously developed Berlin Definition-based inclusion and  
530 exclusion language as a guideline for critical care physicians reviewing chest imaging reports to minimize  
531 interrater disagreements<sup>7</sup>. As a consequence, we are also able to leverage this inclusion and exclusion  
532 language for the NLP processing of chest imaging reports needed for our ML approach. We believe this  
533 choice allowed us to increase the signal-to-noise ratio of the data used for ML development; however, we  
534 recognize that choosing chest imaging reports over the images themselves might limit implementation of  
535 this pipeline for real-time use. On the other hand, we explored the relationship between interrater  
536 disagreement and ML model confidence. We found that lower disagreement rates among our raters  
537 correlated with relatively lower model confidence of “yes”, indicating that our algorithm can confidently  
538 “discard” cases which are not likely to have bilateral infiltrates. Interestingly, we also observed that our  
539 raters exhibited higher levels of disagreement when the model had high model confidence of “yes”. We  
540 speculate this could be attributed to MIMIC III containing more reports that are not consistent with  
541 bilateral infiltrates (88%).

542 Concerning the adjudication of ARDS risk factors, we faced a significant challenge in  
543 implementing machine learning techniques for parsing attending physician notes due to the lack of clearly  
544 defined inclusion/exclusion language for adjudicating ARDS risk factors. In addition, we only had 744  
545 attending physician notes labeled for sepsis, the most of any risk factor, compared to more than 12,000

546 labeled chest imaging reports. Nonetheless, it is striking that even a regex approach for this step yielded  
547 high overall pipeline performance and a small fraction of false negatives.

548         A possible path to improve performance could involve refining the note-parsing process using  
549 advanced natural language processing techniques such as Med-BERT, cTAKES, or leveraging the  
550 capabilities of open-source large language models. The latter technique has the potential to eliminate  
551 laborious pre-processing steps and facilitate the development of general-purpose models instead of task-  
552 specific ones<sup>23</sup>. This is especially true since we use regex patterns for attending physician notes and  
553 echocardiogram reports, which would very likely need redevelopment for each health system in which  
554 our pipeline would get implemented. However, while such advanced approaches offer exciting  
555 opportunities, we must remain cautious as it is in the interest of patients and physicians to implement  
556 approaches that prioritize interpretability and transparency. Not to mention the cost-effectiveness of  
557 developing and deploying pipelines such as ours instead of those relying on large language models (in  
558 token consumption and computational resources, among other costs).

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

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

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

585 **References**

- 586 1. James, J. T. A New, Evidence-based Estimate of Patient Harms Associated with Hospital Care. *J.*  
587 *Patient Saf.* **9**, 122–128 (2013).
- 588 2. Makary, M. A. & Daniel, M. Medical error—the third leading cause of death in the US. *BMJ* i2139  
589 (2016) doi:10.1136/bmj.i2139.
- 590 3. Landrigan, C. P. *et al.* Temporal Trends in Rates of Patient Harm Resulting from Medical Care. *N.*  
591 *Engl. J. Med.* **363**, 2124–2134 (2010).
- 592 4. DeGrave, A. J., Janizek, J. D. & Lee, S.-I. AI for radiographic COVID-19 detection selects shortcuts  
593 over signal. *Nat. Mach. Intell.* **3**, 610–619 (2021).
- 594 5. The ARDS Definition Task Force\*. Acute Respiratory Distress Syndrome: The Berlin Definition.  
595 *JAMA* **307**, 2526–2533 (2012).
- 596 6. Bellani, G. *et al.* Epidemiology, Patterns of Care, and Mortality for Patients With Acute Respiratory  
597 Distress Syndrome in Intensive Care Units in 50 Countries. *JAMA* **315**, 788–800 (2016).
- 598 7. Weiss, C. H. *et al.* Low Tidal Volume Ventilation Use in Acute Respiratory Distress Syndrome\*.  
599 *Crit. Care Med.* **44**, (2016).
- 600 8. Weiss, C. H. *et al.* A Critical Care Clinician Survey Comparing Attitudes and Perceived Barriers to  
601 Low Tidal Volume Ventilation with Actual Practice. *Ann. Am. Thorac. Soc.* **14**, 1682–1689 (2017).
- 602 9. Koenig, H. C. *et al.* Performance of an automated electronic acute lung injury screening system in  
603 intensive care unit patients\*. *Crit. Care Med.* **39**, (2011).
- 604 10. Herasevich, V., Yilmaz, M., Khan, H., Hubmayr, R. D. & Gajic, O. Validation of an electronic  
605 surveillance system for acute lung injury. *Intensive Care Med.* **35**, 1018–1023 (2009).
- 606 11. Laffey, J. G., Pham, T. & Bellani, G. Continued under-recognition of acute respiratory distress  
607 syndrome after the Berlin definition: what is the solution? *Curr. Opin. Crit. Care* **23**, (2017).
- 608 12. Zeiberg, D. *et al.* Machine learning for patient risk stratification for acute respiratory distress  
609 syndrome. *PLOS ONE* **14**, e0214465 (2019).
- 610 13. Song, X., Weister, T. J., Dong, Y., Kashani, K. B. & Kashyap, R. Derivation and Validation of an

- 611 Automated Search Strategy to Retrospectively Identify Acute Respiratory Distress Patients Per Berlin  
612 Definition. *Front. Med.* **8**, (2021).
- 613 14. Li, H. *et al.* Rule-Based Cohort Definitions for Acute Respiratory Distress Syndrome: A Computable  
614 Phenotyping Strategy Based on the Berlin Definition. *Crit. Care Explor.* **3**, e0451 (2021).
- 615 15. Sathe, N. A. *et al.* Evaluating construct validity of computable acute respiratory distress syndrome  
616 definitions in adults hospitalized with COVID-19: an electronic health records based approach. *BMC*  
617 *Pulm. Med.* **23**, 292 (2023).
- 618 16. Johnson, A. E. W. *et al.* MIMIC-III, a freely accessible critical care database. *Sci. Data* **3**, 160035  
619 (2016).
- 620 17. Johnson, A., Pollard, T. & Mark, R. MIMIC-III Clinical Database. [object Object]  
621 <https://doi.org/10.13026/C2XW26> (2015).
- 622 18. Sjoding, M. W. *et al.* Interobserver Reliability of the Berlin ARDS Definition and Strategies to  
623 Improve the Reliability of ARDS Diagnosis. *Chest* **153**, 361–367 (2018).
- 624 19. Chen, T. & Guestrin, C. XGBoost: A Scalable Tree Boosting System. in *Proceedings of the 22nd*  
625 *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 785–794  
626 (Association for Computing Machinery, New York, NY, USA, 2016).  
627 doi:10.1145/2939672.2939785.
- 628 20. Bergstra, J., Yamins, D. & Cox, D. D. Making a Science of Model Search: Hyperparameter  
629 Optimization in Hundreds of Dimensions for Vision Architectures. in *Proceedings of the 30th*  
630 *International Conference on International Conference on Machine Learning - Volume 28 I-115-I-*  
631 *123* (JMLR.org, 2013).
- 632 21. van den Goorbergh, R., van Smeden, M., Timmerman, D. & Van Calster, B. The harm of class  
633 imbalance corrections for risk prediction models: illustration and simulation using logistic regression.  
634 *J. Am. Med. Inform. Assoc.* **29**, 1525–1534 (2022).
- 635 22. Afshar, M. *et al.* A Computable Phenotype for Acute Respiratory Distress Syndrome Using Natural  
636 Language Processing and Machine Learning. *AMIA Annu. Symp. Proc. AMIA Symp.* **2018**, 157–165

- 637 (2018).
- 638 23. Lee, R. Y. *et al.* Assessment of Natural Language Processing of Electronic Health Records to  
639 Measure Goals-of-Care Discussions as a Clinical Trial Outcome. *JAMA Netw. Open* **6**, e231204–  
640 e231204 (2023).
- 641 24. Youden, W. J. Index for rating diagnostic tests. *Cancer* **3**, 32–35 (1950).
- 642 25. Smits, N. A note on Youden’s J and its cost ratio. *BMC Med. Res. Methodol.* **10**, 89 (2010).
- 643 26. Serpa Neto, A. *et al.* Association Between Use of Lung-Protective Ventilation With Lower Tidal  
644 Volumes and Clinical Outcomes Among Patients Without Acute Respiratory Distress Syndrome: A  
645 Meta-analysis. *JAMA* **308**, 1651–1659 (2012).
- 646 27. Determann, R. M. *et al.* Ventilation with lower tidal volumes as compared with conventional tidal  
647 volumes for patients without acute lung injury: a preventive randomized controlled trial. *Crit. Care*  
648 **14**, R1 (2010).
- 649 28. Petrucci, N. & De Feo, C. Lung protective ventilation strategy for the acute respiratory distress  
650 syndrome. *Cochrane Database Syst. Rev.* (2013) doi:10.1002/14651858.CD003844.pub4.
- 651 29. Ancker, J. S. *et al.* Effects of workload, work complexity, and repeated alerts on alert fatigue in a  
652 clinical decision support system. *BMC Med. Inform. Decis. Mak.* **17**, 36 (2017).
- 653 30. Lee, E. K., Wu, T.-L., Senior, T. & Jose, J. Medical Alert Management: A Real-Time Adaptive  
654 Decision Support Tool to Reduce Alert Fatigue. *AMIA. Annu. Symp. Proc.* **2014**, 845–854 (2014).
- 655 31. Cvach Maria. Monitor Alarm Fatigue: An Integrative Review. *Biomed. Instrum. Technol.* **46**, 268–  
656 277 (2012).

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

658

659 **Funding:** FX was supported in part by the National Institutes of Health Training Grant (T32GM008449)

660 through Northwestern University's Biotechnology Training Program; R.A.K.R. CHW was supported by

661 the National Heart Lung and Blood Institute (R01HL140362 and K23HL118139). LANA was supported

662 by the National Heart Lung and Blood Institute (R01HL140362). LANA and FX are supported by the

663 National Institute of Allergy and Infectious Diseases (U19AI135964).

664

665 **Author Contributions:** **FM** - Methodology, Software, Validation, Data Curation, Writing – Original

666 Draft, Writing – Review & Editing, Visualization. **HAL** - Methodology, Software, Validation, Data

667 Curation, Writing – Review & Editing. **HTN** - Software, Validation, Data Curation. **MB** - Methodology,

668 Validation, Data Curation, Writing – Review & Editing. **FX** - Methodology, Software, Validation, Data

669 Curation, Writing – Review & Editing, Visualization. **JK** - Data Curation. **ELC** - Data Curation, Writing

670 – Review & Editing. **CHW** - Conceptualization, Methodology, Validation, Data Curation, Resources,

671 Writing – Original Draft, Writing – Review & Editing, Visualization, Supervision, Project

672 Administration, Funding Acquisition. **LANA** - Conceptualization, Methodology, Software, Validation,

673 Formal Analysis, Resources, Writing – Original Draft, Writing – Review & Editing, Visualization,

674 Supervision, Project Administration, Funding Acquisition.

675

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

677

678 **Data Availability.** The datasets analyzed in this study will be made available upon publication at ARCH

679 repository hosted by Northwestern University (<https://arch.library.northwestern.edu>).

680

681 **Code Availability.** The Python code to reproduce the reported results will be made available upon

682 publication at the Amaral lab GitHub repository (<https://github.com/amarallab>).