Enhancing Extubation Failure Prediction with LLM-Derived Features from Respiratory Therapy Clinical Notes

Izzy Chaiken CHAIKEN@UW.EDU
Information School, University of Washington, Seattle, WA, USA

Aditya Khowal AKHOWAL@UW.EDU
Computer Science & Engineering, University of Washington, Seattle, WA, USA

Neha A. Sathe NAS212@UW.EDU
Department of Medicine, University of Washington, Seattle, WA, USA

Mark M. Wurfel MWURFEL@UW.EDU
Department of Medicine, University of Washington, Seattle, WA, USA

Lucy Lu Wang LUCYLW@UW.EDU
Information School, University of Washington, Seattle, WA, USA

Abstract

Invasive mechanical ventilation is a lifesaving therapy, but timely, safe discontinuation is essential to preventing extubation failure (EF) and related risks to health. We present a novel approach to EF prediction that leverages features classified in free-text respiratory therapy notes using a large language model and logistic regression pipeline. Applied to a patient cohort from University of Washington Medicine, our method identifies clinically meaningful EF-related features that improve EF prediction performance when included alongside structured patient data. We further highlight how differences in target populations in prior EF prediction studies, such as heterogenous inclusion criteria and EF definition, can lead to systematic differences in model performance and hinder generalizability between studies.

Keywords: extubation failure, clinical outcome prediction, large language models, EHR

Data and Code Availability We constructed a novel electronic health records (EHR) dataset from patients at University of Washington Medicine. Our dataset includes adult patients who received invasive mechanical ventilation across 10,810 visits to any of the three hospitals of University of Washington Medicine between April 2021 and September 2023. We utilized both these patients' structured data and unstructured respiratory therapy notes. Personally identifiable information (PII) from our dataset was stored on a HIPAA-compliant server and only accessed by individuals with relevant certification. This dataset has not be made publicly available as it includes PII. Code for reproducing experiments is available at https://github.com/larchlab/extubation-failure-camera-ready.

Institutional Review Board (IRB) This work is approved by the IRB of the Human Subjects Division of University of Washington Medicine under protocol number STUDY00018582.

Author Contributions The first author led data extraction,conducting experiments, analysis of results, and writing this manuscript. The second author aided in designing the LLM note labeling pipeline, labeling clinical notes, and contributing to writing in Section 4. The third and fourth authors are practicing intensive care pulmonologists at University of Washington Medicine; they provided assistance in data collection and interpretation, developing clinical note classification schemas, final labeling of clinical notes, selection of experiments, interpretation of experiment results, and clinical motivation and background included in this manuscript. The final author provided detailed guidance on all aspects of this research.

1. Introduction

Invasive mechanical ventilation (IMV) is a lifesaving therapy for patients with respiratory failure (Wun-

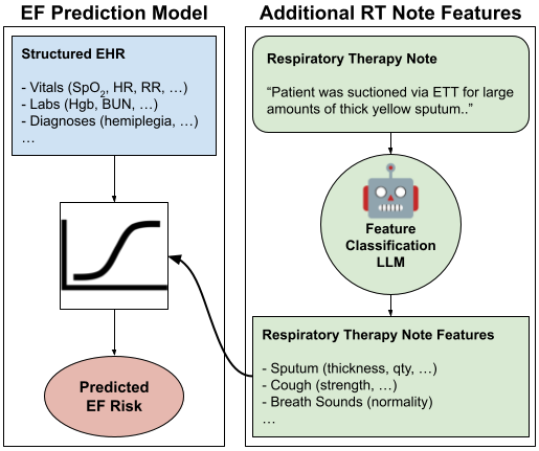


Figure 1: Our EF outcome prediction pipeline. Models developed in prior work include only features derived from tabular EHR (left). Our models include features classified from clinical notes using LLMs to produce more complete risk assessments (right).

sch et al., 2010; Mehta et al., 2015). The decision to extubate (discontinue IMV) is complex and balances the costs of ongoing IMV against the risk of *extubation failure* (EF), when a patient dies or requires reintubation following extubation (Thille et al., 2013). Prolonged IMV is associated with complications such as airway injury, pneumonia, and long-term functional deficits (Herridge et al., 2011, 2003; Jubran et al., 2019; Klompas et al., 2015), while EF itself is associated with prolonged IMV, longer ICU stays, and excess mortality (Thille et al., 2011; Epstein et al., 1997; Béduneau et al., 2017; Frutos-Vivar et al., 2011). Physicians vary in deciding when to extubate and how to use therapies to reduce the risk of extubation failure (Ely et al., 1996; Betbese et al., 1998; Wennberg, 2011).

To support these decisions and help reduce EF, clinicians may use clinical decision support systems (CDSS) to augment and standardize human judgment. These CDSS are typically developed through prospective clinical trials, which directly test whether specific risk factors are related to extubation outcomes (Burns et al., 2025). For example, University of Washington Medicine employs a custom CDSS to assess EF risk, and patients found to be at high risk for EF receive additional evaluation and monitoring by anesthesiologists; in the cohort we examine, roughly half of patients assessed with this tool were classified as high risk and half as low risk, yet these groups had similar EF rates, highlighting inefficiencies in current treatment allocation and opportunities to improve EF prediction accuracy.

Prior research applies machine learning techniques to predict EF, but these models are restricted to using inputs available in tabular electronic health record (EHR) data (e.g. vitals, labs, or ventilator information) (Igarashi et al., 2022). These models do not exploit the rich information in clinical notes collected during IMV, which encode patients' airway and respiratory status. Recent research has demonstrated the feasibility of applying large language models (LLMs) to generate outcome predictions either directly from unstructured notes (Van Aken et al., 2021; Naik et al., 2022), or to extract entities from notes at scale for incorporation into predictive models (Robitschek et al., 2025; Mugisha and Paik, 2022). Building on clinical information extraction and EF prediction literature, we define and design a novel LLM-based pipeline to classify features in respiratory therapy (RT) notes.[1] We develop EF prediction models including these features, and demonstrate that they improve prediction performance, revealing new risk factors that are important to document and model.

Prior EF research also uses inconsistent definitions for inclusion criteria. Figure 2 illustrates an example clinical time course for a patient who experiences EF: an initial IMV episode begins at time T_1 and ends with extubation at T_2, before the patient experiences EF at T_3 when IMV is reapplied. Patient cohorts for EF risk investigations use a *minimum IMV duration* ($T_2 - T_1$) to define valid initial IMV episodes for inclusion. Failure events are then labeled according to a *maximum failure window*—EF occurs if failure is within this window but not after (i.e., if $T_3 - T_2$ is greater than this window, the patient will be labeled negative for EF, despite experiencing reintubation). Patients with different initial IMV durations exhibit variable EF risk profiles (Thille et al., 2019), impeding comparability of models trained using different cohort definitions and hindering consistent interpretations of EF risk factors (Nava et al., 2005). Table 1 shows this inconsistency in definitions adopted by prior work. For example, most modeling studies define EF as occurring within 2-3 days of extubation (Torrini et al., 2021), whereas prospective clinical studies instead use a 7-day window (Béduneau

1. These notes are distinct from those found in publicly available EHR datasets such as MIMIC-IV (Johnson et al., 2023), as they include descriptions of a patient's state collected during IMV.

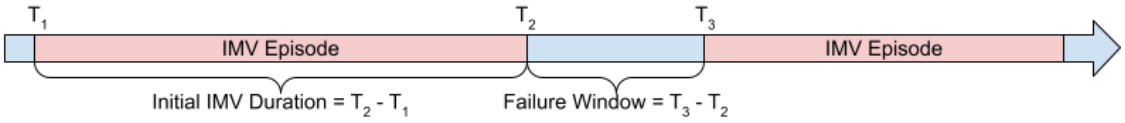


Figure 2: Example clinical time course for patient experiencing EF. Minimum initial IMV duration and maximum time to failure (i.e., the *failure window*) are used to define cohorts for modeling and whether a patient experiences EF. We assess the effects of varying these two criteria on downstream performance. Because of the frequency of variable collection, structured model features from the 4 hours prior to the end of the initial IMV episode and note features from 12 hours prior to the end of the initial IMV episode are included in the EF prediction models.

et al., 2017). To focus our modeling on high risk patients and ensure all clinically relevant EF is labeled as such, we estimate EF risk among patients with IMV duration of at least 24 hours and define failure as occurring within 7 days of extubation. We further train models with varying EF definitions and inclusion criteria and show their impacts on model performance and interpretation.

Our contributions are as follows:

- We introduce a schema of 15 features described in unstructured respiratory therapy (RT) notes that are relevant to EF, e.g., cough strength and sputum quantity. We develop LLM-based classifiers that reliably identify presence of these features; we validate against a set of 200 RT notes manually annotated for feature values (§4).
- We construct a cohort of 10k+ patients who underwent IMV at University of Washington Medicine (§3), and incorporate their RT note features into EF risk models (§5). We find that RT note features improve AUROC by 1.9 points for our best performing models, with a larger difference when models are trained only over patients with documented RT notes (5.1 points) (§6).
- We investigate how different inclusion criteria and EF definitions impact model performance (§6). Varying minimum IMV duration strongly impacts performance metrics: AUROC generally decreases and F1 increases as minimum IMV duration is increased. Since performance of EF predictors varies with characteristics of the patient population, models trained on cohorts with different inclusion criteria are not directly comparable.

2. Related Work

Clinical Assessment of EF Risk Prior studies have developed tools to estimate patient-level EF risk, which guide clinical decisions at the time of planned extubation. If a patient is deemed high-risk for EF, a clinician may defer extubation and/or initiate treatments to reduce risk of failure (Thille et al., 2019; Apfelbaum et al., 2021; Quintard et al., 2019; Grieco et al., 2021). However, the lack of standardized EF risk assessments leads to variation in assessed risk levels and choice of extubation-time therapies (Burns et al., 2018; Godard et al., 2016). A variety of risk-stratification systems for EF are used in clinical practice (Sarti et al., 2021; Joffe and Barnes, 2022), integrating features such as real-time vitals, medical history, and subjective impression of risk (Burns et al., 2025; Joffe and Barnes, 2022). Such systems improve patient outcomes during extubation, and similar systems for other aspects of hospital care improve adherence to care guidelines and reduce morbidity (Zheng et al., 2022; Moja et al., 2014).

Machine Learning for EF Prediction Recent research develops machine learning models to retrospectively predict EF risk in adult ICU patients using structured variables (Chen et al., 2019; Fabregat et al., 2021; Fleuren et al., 2021; Hsieh et al., 2018; Otaguro et al., 2021; Seely et al., 2014; Zhao et al., 2021). Some models also include longitudinal data (Seely et al., 2014; Zeng et al., 2022). Prior EF prediction models largely utilize gradient boosting (GB) and artificial neural networks (ANN) (Igarashi et al., 2022), and attain AUROC between 0.83-0.85 on variable-sized cohorts (Zhao et al., 2021; Hsieh et al., 2018). However, EF prediction performance may not generalize to external cohorts: Zhao et al. (2021) demonstrated their model performance diminishing when tested on a patient cohort from a hospital not included in the training data. Models from prior work are generally unavailable, limiting our ability to assess their performance over our cohort.

EF Definition & Selection Criteria EF is not consistently defined, as the window during which EF

Source	Model Type	Cohort Size	EF Rate	Failure Window (days)	Min Intub Len (hrs)
Seely et al. (2014)	LR ensemble	434	12%	2	48
Hsieh et al. (2018)	ANN	3602	5%	3	0
Chen et al. (2019)	GB	3636	17%	2	0
Fabregat et al. (2021)	SVM	1108	9%	7	12
Fleuren et al. (2021)	GB	883	19%	7	24
Otaguro et al. (2021)	GB	117	11%	3	24
Zhao et al. (2021)	GB	16189	17%	2	0
Zeng et al. (2022)	RNN	8599	30%	2	12
Ours	LR	3243	20%	7	24

Table 1: Model type, cohort statistics, and inclusion criteria reported in prior literature. EF rate contextualizes the overall outcomes within a patient cohort, providing another indicator of variability within experimental settings

may occur varies. Therapies such as non-invasive ventilation can delay EF, so shorter EF windows may improperly label patients (Thille et al., 2016). No standard minimum initial IMV duration in used in research, so EF risk factors and outcome predictions may be inconsistent (Rose et al., 2017). IMV duration is itself associated with higher risk of EF, both due to differences in patient population at various durations, and through physiologic effects of prolonged IMV (Rothaar and Epstein, 2003; Torrini et al., 2021). Clinical literature uses IMV duration as a proxy for EF risk, so prior studies differ in patient risk profiles (Nava et al., 2005; Thille et al., 2013). We examine the effects of EF failure windows and cohort selection criteria on EF risk models.

Clinical Note Feature Classification LLMs have demonstrated high performance in determining information contained in biomedical texts (Perera et al., 2020; Li et al., 2024), enabling information extraction from clinical notes at scale using few-shot learning techniques (Agrawal et al., 2022; Goel et al., 2023). LLMs have been applied to extract features from clinical notes to predict outcomes such as contraceptive switching rationales (Miao et al., 2025), bladder cancer survival (Sun et al., 2024), concepts related to postpartum hemorrhage (Alsentzer et al., 2023), and breast cancer phenotypes (Zhou et al., 2022). We extend this methodology, classifying features related to a patient's IMV status in RT notes.

3. Data

Cohort As described above, we assemble a cohort of 10,194 patients who received IMV across 10,810 visits to any of the three hospitals of University of Washington Medicine between April 2021 and September 2023. We define *extubation* as the time at which a patient's documented oxygen delivery device changes from IMV to a non-IMV method of delivery, and *extubation failure* as patient death or a return to IMV within 7 days of initial extubation.

Of the initial 10,810 encounters, we drop 843 for having incomplete height/weight information, 4,711 for having no IMV session lasting at least 24 hours, 1,922 for having a 'Do Not Intubate/Resuscitate' order within 7 days of extubation, and 91 for being encounters with a non-unique patient. Our primary dataset includes 3,243 IMV sessions from unique patients, of whom 647 (19.95%) experienced EF within 7 days; 618 of these are due to reintubation, and 29 are due to death. Among patients in the non-excluded cohort, median IMV duration was 60.8 hours, and mean IMV duration was 97.1 hours. The first and third quartiles were 36.6 hours and 112.0 hours. Only 2,339 encounters have corresponding readiness checklists (our clinic's CDSS tool, which queries for the presence of 19 binary factors to assess EF risk) from within 4 hours of extubation. We hold out 646 (20%) encounters as a test set, based on a random sample stratified by documented patient race/ethnicity and EF outcome.

To investigate the impact of inclusion criteria, we further include 3,685 encounters from patients not

Category	Feature	Macro Rec/Prec/F1
Sputum	Presence	0.933/0.943/0.937
	Thick	0.988/0.996/0.992
	Thin	0.944/0.997/0.969
	Quantity	0.881/0.852/0.857
	Color	0.876/0.808/0.816
Cough	Presence	0.995/0.989/0.991
	Weak	0.780/0.825/0.801
	Strong	0.935/0.923/0.929
	Induced	0.835/0.783/0.806
	Spontaneous*	0.888/0.545/0.520
Suctioning	Presence*	0.875/0.872/0.873
	Oral*	0.713/0.834/0.756
	Endotracheal*	0.709/0.798/0.732
Cuff Leak	Presence	0.991/0.940/0.963
Breath Sounds	Normality	0.929/0.920/0.924

Table 2: Metrics for features extracted from RT Notes. *variables not included in downstream models due to low F1 or high correlation with other features.

in the aforementioned set whose initial intubation was at least 1 hour, rather than 24. Of these additional patients, 133 (3.61%) experienced EF. We retain 741 of these patients for our test set, based on the same stratified sampling strategy. See Appendix 5 for dataset demographics.

4. RT Note Feature Classification

Respiratory therapists support clinicians in preventing and treating respiratory diseases. They regularly assess clinical features that would necessitate change in respiratory support and document their findings in free text respiratory therapy (RT) notes. Our dataset includes 34,834 RT notes collected before each patient's initial extubation, and we hypothesize that features documented in these notes may be valuable predictors for extubation failure. Relevant features include cough strength and secretion quantity, which both reflect the ability for patients to clear their airways, and are well-established risk factors for extubation failure (Ouanes-Besbes et al., 2012; Chien et al., 2008; Mekontso-Dessap et al., 2006; Duan et al., 2021).

RT note excerpt:

... No vent changes. Breath sounds clear. Suctioning for small amounts of thick white secretions. Strong cough, +gag, +cuff leak. Plan SBT and extubate after MRI ETT 7 27 @ the teeth ~5^ AMV RR 14, Vt 500, +5 ...

Labels:
Sputum: Present=1, Consistency=1, Quantity=1, Color=0
Cough: Present=1, Strength=1, Induced=0
Breath Sounds: 0
Cuff Leak: 1

Figure 3: Excerpt of a respiratory therapy note and feature labels in our dataset. Spans corresponding to feature values are highlighted.

We define 15 features related to five categories, specified in Table 2, based on a review of literature and pulmonologist input. We then define a novel feature classification task to classify these possible predictors for EF in RT notes. For the EF risk assessment task, we are interested in patient-level qualities, so we prompt LLMs to classify entire RT notes, rather than named entity recognition to identify specific note spans. We investigate few-shot prompting due to the lack of a large-scale labeled dataset.

Feature Annotation To support prompt engineering and evaluate classification performance, we manually label feature values for a random sample of 400 total RT notes (200 from the training split and 200 from test). We develop annotations collaboratively over multiple rounds of review. The first and second authors label and come to a consensus on labels for each RT note, and these annotations are reviewed and corrected by two practicing critical care pulmonologists. Any disagreements are discussed and a consensus decision is made. An excerpt from an example labeled note is shown in Figure 3. We perform prompt engineering exclusively using 200 RT notes from the training split. We report final performance metrics for the classification tasks over the 200 RT notes from the test split.

Few-shot Prompting We prompt Meta-Llama-3-8B-Instruct (Touvron et al., 2023) to classify feature values. We sample excerpts from the labeled notes of the training split to represent each value of each feature for use as in-context learning examples (Brown et al., 2020). We design separate prompts for suctioning, sputum, cough, breath sounds, and cuff

Category	Feature	Downstream Feature Mapping	N Mentions
Sputum	Present	{Unlabeled: 0, Positive: 1}	1734 (53.47%)
	Consistency	{Unlabeled: 0, Thin: -1, Thick: 1}	748.0 (23.07%)
	Quantity	{Unlabeled: 0, Low: 1, Medium: 2, High: 3}	1478 (45.58%)
	Color**	{Unlabeled: 0, Non-Pathological: 0, Pathological: 1}	305.0 (9.40%)
Cough	Present	{Unlabeled: 0, Negative: -1, Positive: 1}	676 (20.84%)
	Strength	{Unlabeled: 0, Weak: -1, Strong: 1}	360 (11.10%)
	Induced	{Unlabeled: 0, Induced: 1}	128 (3.95%)
Breath Sounds	Normality	{Unlabeled: 0, Normal: 0, Abnormal: 1}	1892 (58.34%)
Cuff Leak	Presence	{Unlabeled: 0, Absent: 0, Present: 1}	152 (4.69%)

Table 3: Encodings of RT note features included in EF risk models, and the number of patients for whom our LLM pipeline classified any mention of this feature. **Non-pathological sputum colors include clear, white, and tan; pathological sputum colors include yellow, green, pink, red, and rust.

leak features; and prompt the LLM to respond with predefined answers such as *yes/no* for binary features or explicit categorical labels (e.g., *low/medium/high* for sputum amount). All features may be classified as *Unlabeled*, indicating that no explicit feature value is mentioned in the RT note. Temperature was set to 0.01 for all extractions. Final prompts can be found in the supplemental code.

RT Note Classification Performance The mean tokenized RT note length is 258.26 (min 3, max 3178). Over the test set, Macro-F1 for each variable ranges from 0.53 (for spontaneous cough) to 0.99 (for thick sputum), although most F1 values are above 0.80 (see Table 2).

Classification Results We run our final feature classification pipeline over the entire corpus, on RT notes in the 12 hours proceeding extubation events. We have 2,869 such notes from 2,509 of 3,243 valid encounters in our dataset (distribution of notes per patient available in Appendix A). Some features are present at high rates (abnormal breath sounds at 58.3%), while others are rare (absent cough at 1.2%). Patients with RT notes classified as having high sputum quantity and weak cough co-occur with EF at higher rates than the overall study population, indicating potential association between these features and EF outcomes.

5. EF Prediction Methods

We train logistic regression (LR) and gradient boosting (GB) models to predict, for each patient, their binary outcome of extubation failure (0=success, 1=failure).

Structured EHR Predictor Variables For each patient, we collect the following structured EHR variables (part of the *baseline* input variable set): IMV episode duration, vitals (e.g. SpO_2, heart rate, temperature), labs (e.g. hemoglobin, creatinine, calcium), ventilation data (e.g. AutoPEEP, FiO_2, tidal volume), diagnoses at admission time (e.g. acute respiratory failure, myocardial infarction, chronic pulmonary disease), and demographics of age and documented sex. We also compute medication doses adjusted for a patient's body weight (e.g. opioid doses, vasopressors, and propofol), which we include as the *medication* input variable set. Since there are multiple measurements, we use the average value of vital, lab, medication, and ventilation features in the 4-hour window before extubation. Appendix E contains a full list of predictors.

Input Variables We train predictive models with the following combinations of input variables:

- Baseline (B): structured EHR variables such as length of initial intubation, vitals, labs, ventilation data, admission-time diagnoses, age, and documented sex;
- Medication (M): normalized medication doses in the 4 hours before extubation; we include medica-

tions that are commonly delivered to patients while on IMV, which potentially have negative effects after extubation;
- RT Note (N): features classified in respiratory therapy notes as described in §4. We omit suctioning and spontaneous cough variables because of lower extraction performance (F1<0.8) or high correlation with other RT note features (e.g., suctioning presence is highly correlated with sputum presence ($r = 0.808$)). Variable values are encoded as described in Table 3.

Model Variants For LR and GB models, we train and test four model variants: one with only base variables (e.g., LR_B), one with base and medication variables (e.g., LR_{B+M}), one with base and RT note variables (e.g., LR_{B+N}) and one with base, medication, and RT note variables (e.g., LR_{B+M+N}). These variants allow us to assess the impacts of including each set of variables not included in prior work. In supplementary analysis, to assess the temporal generalizability of our models we also fit variants of the LR_{B+M+N} models on the subset of patients admitted during 2021-22, and report performance on a test set consisting of patients admitted in 2023 (results in Appendix H).

We also assess whether survival analysis models are more effective for estimating probable time to extubation failure. Rather than modeling the probability of a single binary outcome, these models estimate the relationship between predictor variables and the probability a patient will not experience a particular outcome, in our case EF, at a particular time. We assess whether such models are better at predicting EF by accounting for the temporal information of when specific failures occurred or if a patient was discharged. Specifically, we train Cox Proportional Hazards and gradient boosting survival analysis models, fit over the complete set of base, medication, and note variables. The outcome variable of these models is time from extubation to EF, and we censor patients at discharge time. We binarize the outputs of these models by assessing the estimated probability of survival (i.e, not experiencing extubation failure) after seven days.

Model Evaluation & Analysis We report AUROC on the held-out test set as our primary metric. Intubated patients form a diverse cohort both in terms of medical conditions and demographic attributes, so we select AUROC because it reflects both positive and negative class performance (McDermott et al., 2024). We additionally report AUPRC, Sensitivity (Recall), Specificity, PPV (Precision), NPV, F1, and Accuracy. Where appropriate, the threshold for patients deemed high risk is selected based on the prevalence of EF in the training split. We report features with high magnitude coefficients in our LR models (all features are normalized before training).

We further contextualize the performance of our models with the risk predictions made by the readiness checklist CDSS used at University of Washington Medicine. The checklist includes 19 yes/no questions; per local guidelines, patients with ≥2 positive responses are deemed high risk for EF and managed with enhanced post-extubation monitoring. Appendix C contains the full list of features in this checklist.

Varying Inclusion Criteria and EF Definition To study the impacts of alternate criteria, we train variants of our best performing model: LR_{B+M+N}, while varying the inclusion criteria and EF definition.
- Minimum IMV duration: we vary the minimum IMV duration from 1 hour to 24 hours, while holding the size of the training split constant. For each model variant, we test on a held out split of 20% of patients meeting the same criteria as the train set. Test split size varies, though any encounter that is used in the test set for any experiment is never used in the training sets for any experiment.
- Failure window: we vary the maximum EF window from within 12 hours of extubation to within 336 hours (14 days). We use constant train/test splits differing only by outcome labels, so EF rate ranges from 8.24% after 12 hours to 22.72% after 14 days.

Implementation We implement all LR and GB models using `scikit-Learn` (Pedregosa et al., 2011). We trained survival analysis models using the `scikit-survival` library (Pölsterl, 2020). We perform cross-validation over the train split to find hyperparameters maximizing AUROC. Final results are reported on the held-out test set. Further details in Appendix D.

6. EF Prediction Results

RT note variables improve EF prediction Logistic regression models incorporating RT note features demonstrate the strongest performance on our cohort, with consistent gains across most metrics we report (Table 4). For example, adding RT note features to the baseline logistic regression model in-

Model & Features	Cohort	AUROC	AUPRC	Sens. (Recall)	Spec.	PPV (Prec.)	NPV	$F1_+$	$F1_-$	Acc.
Checklist**		—	—	0.55	0.55	0.21	0.84	0.31	0.66	0.55
LR_B	All	0.729	0.380	0.677	0.663	0.323	0.896	0.438	0.762	0.666
LR_{B+M}		0.733	0.388	0.694	**0.670**	0.333	0.902	0.450	**0.769**	0.675
LR_{B+N}		0.749	0.392	**0.750**	0.657	**0.342**	**0.917**	**0.470**	0.766	0.675
LR_{B+M+N}		**0.752**	**0.399**	0.734	0.665	**0.342**	0.913	0.467	**0.769**	**0.678**
GB_B	All	0.671	0.308	0.597	0.665	0.297	0.874	0.397	0.755	0.652
GB_{B+M}		0.671	0.307	0.597	0.657	0.292	0.872	0.393	0.750	0.646
GB_{B+N}		0.669	0.309	0.589	0.661	0.292	0.871	0.390	0.752	0.647
GB_{B+M+N}		0.671	0.308	0.573	0.661	0.286	0.867	0.382	0.750	0.644
LR_B	Patients	0.696	0.386	0.680	0.633	0.330	0.881	0.444	0.737	0.643
LR_{B+M}	with RT	0.699	0.393	0.660	0.615	0.313	0.872	0.425	0.721	0.624
LR_{B+N}	notes	0.715	0.403	0.699	0.618	0.327	0.885	0.445	0.728	0.635
LR_{B+M+N}		**0.750**	**0.405**	**0.702**	**0.670**	**0.336**	**0.904**	**0.454**	**0.770**	**0.676**
Survival Analysis Models										
CPH_{B+M+N}	All	0.742	0.394	0.710	0.649	0.325	0.904	0.446	0.756	0.661
GBS_{B+M+N}		0.677	0.336	0.677	0.565	0.270	0.881	0.386	0.689	0.587

Table 4: Logistic regression (LR) and Gradient boosting (GB) EF prediction models with different input features (B: structured EHR features; M: medications; N: features from RT notes) for all patients and only patients with RT notes. We report results for survival analysis methods (Cox Proportional Hazard (CPH) and Gradient Boosting survival analysis models (GBS) when using all input features. **Bold** indicates best metric value over that cohort; all cohorts defined as patients with 24-hour minimum intubation time. Thresholded metrics are computed at the proportion of positives in the training split (0.201).

**We also show metrics from the existing checklist CDSS to contextualize our results, though model and checklist metrics are not directly comparable since high risk as classified by the checklist is used to provision additional treatment to reduce EF risk.

creases AUROC from 0.729 (LR_B; 95% CI: [0.684, 0.775]) to 0.749 (LR_{B+N}; 95% CI: [0.701, 0.795]). Likewise, adding note variables to the model including base and medication inputs increased performance from 0.733 (LR_{B+M}; 95% CI: [0.684, 0.783]) to 0.752 (LR_{B+M+N}; 95% CI: [0.703, 0.794]). The performance gain when adding RT note variables is even greater among the subset of patients with RT notes within 12 hours of extubation. Appendix G describes performance within demographic subgroups; while metric values varied, no performance differences across sex, age, and racial/ethnic subgroups were found to be statistically significant.

Prior EF prediction research tended to find that GB models outperformed LR models (Chen et al., 2019; Fleuren et al., 2021; Otaguro et al., 2021; Zhao et al., 2021). We hypothesize that performance differences may be related to overfitting; to assess this hypothesis we measure model performance over the training set. Over our *training* data, each GB model performed better than the LR model trained using the same inputs. For example, the LR model including all variables attained an AUROC of 0.712 over the training set, whereas the similar GB model attained an AUROC of 0.782 over the training set, indicating that the more complex GB models may not be as robust to distribution shifts between the train and test sets as the LR models.

The Cox Proportional Hazards model fit over all variables attained an AUROC of 0.742 (95% CI: [0.691, 0.790]). The gradient boosting survival analysis model fit over all variables attained an AUROC of

0.677 (95% CI: [0.625, 0.732]). These results suggest that for the binary EF prediction task, naive application of survival analysis models may not generate additional predictive performance.

IMV duration is consistently the most important feature in each model (a LR model with only IMV duration as input feature attains an AUROC of 0.658). Other important variables in the LR model by coefficient magnitude include ventilator plateau pressure, SpO_2, urea nitrogen, and presence of diffuse traumatic brain injury (ICD-10 code S06.2). Sputum quantity and thickness as classified in RT notes are also among the 20 most important predictors.

Our model demonstrates improvement over the existing checklist CDSS. A total of 81 patients in our test set assessed using the CDSS experienced EF, yet the CDSS classifies only 40 of these patients as high risk. Our LR_{B+M+N} model correctly classifies 59 of these 81 patients as high risk. On the other hand, 32 of 40 patients assessed to be high risk by the CDSS are found to be high risk by our LR model, so the two models remain complementary. The CDSS attains 21% precision over patients who experienced EF, while our best model achieves 34%; although some proportion of patients predicted to be high risk by the CDSS may have avoided EF because clinicians acted upon these high risk assessments.

We conduct calibration analysis comparing logistic regression models with and without RT note features (Appendix I). While neither the LR_{B+M} nor the LR_{B+M+N} models are perfectly calibrated (both have calibration curve slopes over 1), they exhibit similar calibration performance.

Increasing minimum IMV duration induces performance trade-offs Model performance varies systematically with inclusion criteria. We observe that including patients with shorter duration IMV when training results in a model that more easily classifies positive patients with longer IMV and negative patients with shorter IMV: the model with a 1-hour minimum IMV duration attains recall of 0.258 and specificity of 0.931 for patients whose IMV duration was <24h, as compared with recall of 0.855 and specificity of 0.489 for those whose IMV was ≥24h. Correspondingly, AUROC drops from 0.796→0.752 when varying minimum IMV duration from 1 to 24 hours (Figure 4(*a*)). Positive class F1 increases (0.390→0.467), driven by higher precision (0.265→0.342) with stable overall recall (0.735→0.734), while negative class F1 decreases (0.795→0.777).[2] GB models exhibited similar AUROC and positive F1 trends, but did not consistently vary in negative class F1 (Appendix F).

Changes in predictor importance between the 1-hour and 24-hour minimum IMV duration cohorts indicate possible cohort-level differences. While IMV episode length, SpO_2, urea nitrogen, plateau pressure, and hemoglobin are the top four predictors in each model, Acute Respiratory Failure, age>60, and FiO_2 are among the top predictors for the 24-hour minimum model, but not the 1-hour minimum model (Appendix Table 7). Additionally, characteristics of these groups vary: for instance, 12.7% of patients who were intubated <24 hours were in a surgical unit for their entire intubation, as compared with 5.2% of patients who were intubated ≥24 hours (p<0.001).

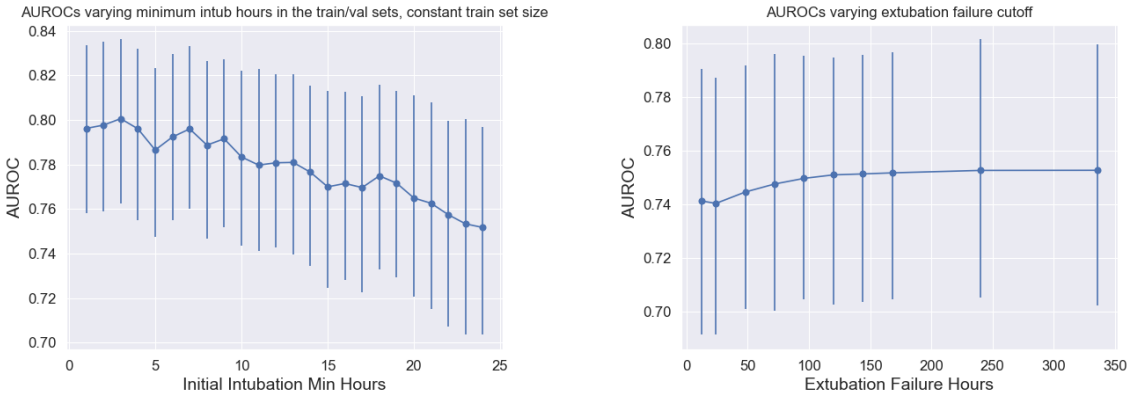
Varying EF window does not systematically alter metrics When increasing maximum EF window from 12 to 336 hours, we do not observe major changes in overall predictor performance per AUROC (Figure 4(*b*)) or positive/negative class F1 (Appendix Figure 10 shows changes in F1), despite the increase in EF prevalence from 8% to >21%. IMV duration remains the most important predictive feature across different choices for EF window.

7. Discussion & Conclusion

Our pipeline classifies a novel set of variables in free-text respiratory therapy notes, a type of clinical note that as far as we know, has not been studied in prior clinical LLM work. LLMs demonstrate strong performance in classifying these features, especially those related to sputum, breath sounds, cuff leak presence, and cough presence and strength.

Predictive models including RT note features perform better than models without these features. These increases remain consistent to robustness checks such as inclusion of medication variables in the model, and disappear when shuffling the columns corresponding to the RT note features. These results imply clinically relevant improvements in predictive performance: Krinsley et al. (2012) suggest 5% may be a reasonable target extubation failure rate (i.e, only 5% of the population are false negatives for detection of extubation failure). To attain this extubation failure rate over our test data, a model must have sensitivity of 74.2%. At this sensitivity, the

2. In our experiments, the most permissive minimum IMV duration (with same train set size) results in highest AUROC (≈0.80), similar to metrics reported in prior work.



(*a*) AUROC varying minimum IMV length (*b*) AUROC varying extubation failure cutoff

Figure 4: Change in AUROC when varying minimum IMV duration and EF window. AUROC tends to decrease as the minimum IMV duration increases (left); AUROC is unchanged as the maximum EF window increases (right). In both experiments, the proportion of patients who experienced EF increases: from 11.26% to 19.95% in the case of inclusion variation, and from 8.24% to 22.72% in the case of EF definition variation. Each different minimum IMV duration model is trained over varying sets of patients with fixed training split size; training data for each EF window model is identical.

base LR model has a FPR of 40.0%, while the LR model augmented with RT note features has a FPR of 34.1%. This difference in FPR corresponds to 5 fewer false alarms per 100 patients distributed identically to our test set, preventing harmful impacts of unnecessary continuation of IMV. A corresponding decision curve analysis indicates that in settings where clinicians target roughly one true case of extubation failure per five false alarms, the additional RT note features may yield one additional true positive per 100 patients treated. Appendix J contains further decision curve analysis details.

We identify high volume of sputum, thick sputum, and weak cough as associated with higher EF risk, so these features may be valuable to document as structured EHR elements in the future. Meanwhile, other significant features in our model, such as long IMV duration, low SpO_2, blood urea nitrogen, and brain injury are known to be associated with extubation failure, providing external validity for our results (Igarashi et al., 2022; Vidotto et al., 2008).

While prior work reports that gradient boosting variants outperformed logistic regression for EF prediction, our work demonstrates that this may not apply to every EF prediction setting. As Christodoulou et al. (2019) show, other machine learning methods do not exhibit significant improvements over logistic regression for clinical prediction models, so there is precedent for such findings. We additionally find that models developed using survival analysis objectives did not exhibit significantly improved performance on binary extubation failure prediction tasks.

Models trained and assessed using different inclusion criteria exhibit performance differences, indicating a potential threat to model generalizability. In our results, AUROC decreases and F1 increases as the minimum IMV duration threshold is raised. Higher IMV duration is associated with higher EF risk, and setting a higher threshold tends to remove more-easily classified lower-risk patients from the population, leading to lower AUROC (McDermott et al., 2024). The relatively poor performance of models greater minimum intubation durations reflect the difficulty of distinguishing risk among patients where the task is clinically meaningful: clinicians are less likely to need risk assessment support for patients with shorter, post-operative intubations. This highlights a broader implication: collaboration with clinicians is essential to ensure that model tasks can be specified in relation to meaningful clinical needs.

Our analysis also yields some counterintuitive observations, such as diagnosis of acute respiratory failure (ARF) being associated with *lower* odds of EF. It is possible that ARF may not be documented as a diagnosis when comorbid with other IMV causes such as head injury or stroke. These diagnoses may induce

treatment differences, possibly altering EF risk. Due to ours and similar studies being retrospective, without counterfactuals, it is difficult to disentangle these mechanisms. We leave work that controls for diagnosis and treatment effects to future clinical studies.

In this work, a key advantage of using classified feature values versus dense note embeddings is interpretability in the downstream model (at the cost of less expressivity). Future work could investigate how best to preserve additional details from RT notes and other unstructured notes in downstream models. Another future direction is investigating how alternate modeling methods, such as time series models, may improve EF prediction, especially when incorporating RT note features. Lastly, EF risk factors may also be associated with demographic groups; e.g., Thille et al. (2023) uncover sex-related differences in IMV and EF. Future work should investigate whether the features we uncover are broadly useful or are only associated with increased EF risk in specific demographic or clinical groups.

Limitations Due to dataset constraints, we are limited to studying EF in a single hospital system located in a large US city. No comparable dataset is publicly available (other public EHR datasets do not contain RT notes), impeding external validation of our results. Both our LLM feature classification pipeline and the downstream EF prediction models may fail to generalize to other clinical notes and patient populations, other languages besides English, or even to the same hospitals over time as the population of intubated patients can change.

We lack treatment counterfactuals: outcomes may have differed had patients not been treated as high risk in the clinical setting, so we cannot assess causal associations between patient features and extubation outcomes. The decision to extubate can also be influenced by non-clinical factors such as family wishes or transitions to palliative care. While DNR/DNI are the most objective way of excluding patients ineligible for re-intubation, future work should investigate whether relevant non-clinical factors can be derived from unstructured notes and quantify their impacts on extubation/re-intubation decisions and outcomes. A larger sample of manually labeled notes may be useful for future work aiming to develop a more robust pipeline.

Conclusion We demonstrate that LLMs in few-shot settings can classify features relevant to extubation failure in free text clinical notes collected during IMV. These features improve performance of downstream EF risk models, enabling identification of additional risk factors that may be useful to include in future CDSS. Furthermore, we describe how inconsistent cohort inclusion criteria are prevalent in related work yet drive changes in model performance, demonstrating a threat to generalizability beyond differences across hospital settings. This result reveals the need for standardized task definitions to enable model comparability and support translation of risk prediction models into clinical practice.

Acknowledgements

This work was supported by the University of Washington Institute of Medical Data Science Pilot Award, the eScience Institute's Cloud Credits for Research and Teaching Program, and gift funds from the Allen Institute for AI. The first author was partially supported by the National Science Foundation CS-Grad4US Fellowship.

References

Monica Agrawal, Stefan Hegselmann, Hunter Lang, Yoon Kim, and David Sontag. Large language models are few-shot clinical information extractors. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1998–2022, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics.

Emily Alsentzer, Matthew J Rasmussen, Romy Fontoura, Alexis L Cull, Brett Beaulieu-Jones, Kathryn J Gray, David W Bates, and Vesela P Kovacheva. Zero-shot interpretable phenotyping of postpartum hemorrhage using large language models. *NPJ digital medicine*, 6(1):212, 2023.

Jeffrey L Apfelbaum, Carin A Hagberg, Richard T Connis, Basem B Abdelmalak, Madhulika Agarkar, Richard P Dutton, John E Fiadjoe, Robert Greif, P Allan Klock, David Mercier, et al. 2022 american society of anesthesiologists practice guidelines for management of the difficult airway. *Anesthesiology*, 136(1):31–81, 2021.

C Heather Ashton. Benzodiazepines: How they work and how to withdraw. *The Ashton Manual, Aug*, 2002.

Gaëtan Béduneau, Tai Pham, Frederique Schortgen, Lise Piquilloud, Elie Zogheib, Maud Jonas, Fabien Grelon, Isabelle Runge, Nicolas Terzi, Steven Grange, et al. Epidemiology of weaning outcome according to a new definition. the wind study. *American journal of respiratory and critical care medicine*, 195(6):772–783, 2017.

Antoni-Jordi Betbese, Manuel Perez, Ela Bak, Gemma Rialp, and Jordi Mancebo. A prospective study of unplanned endotracheal extubation in intensive care unit patients. *Critical care medicine*, 26(7):1180–1186, 1998.

Kathleen Broglio and Russell K Portenoy. Approximate opioid dose conversions and oral total daily morphine milligram equivalents (mmes; refer to important notes below), 2022.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

Karen EA Burns, Stavroula Raptis, Rosane Nisenbaum, Leena Rizvi, Andrew Jones, Jyoti Bakshi, Wylie Tan, Aleksander Meret, Deborah J Cook, Francois Lellouche, et al. International practice variation in weaning critically ill adults from invasive mechanical ventilation. *Annals of the American Thoracic Society*, 15(4):494–502, 2018.

Karen EA Burns, Jill E Allan, Emma Lee, Marlene Santos-Taylor, Phyllis Kay, Pamela Greco, Hilary Every, Owen Mooney, Maged Tanios, Edmund Tan, et al. Liberation from mechanical ventilation using extubation advisor decision support (leads): protocol for a multicentre pilot trial. *BMJ open*, 15(3):e093853, 2025.

Tingting Chen, Jun Xu, Haochao Ying, Xiaojun Chen, Ruiwei Feng, Xueling Fang, Honghao Gao, and Jian Wu. Prediction of extubation failure for intensive care unit patients using light gradient boosting machine. *IEEE Access*, 7:150960–150968, 2019.

Jung-Yien Chien, Mao-Shin Lin, Yuh-Chin T Huang, Yu-Fen Chien, Chong-Jen Yu, and Pan-Chyr Yang. Changes in b-type natriuretic peptide improve weaning outcome predicted by spontaneous breathing trial. *Critical care medicine*, 36(5):1421–1426, 2008.

Evangelia Christodoulou, Jie Ma, Gary Stephen Collins, Ewout Willem Steyerberg, Jan Yvan Jos Verbakel, and B. Van calster. A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. *Journal of clinical epidemiology*, 110:12–22, 2019.

Jun Duan, Xiaofang Zhang, and Jianping Song. Predictive power of extubation failure diagnosed by cough strength: a systematic review and meta-analysis. *Critical Care*, 25(1):357, 2021.

E Wesley Ely, Albert M Baker, Donnie P Dunagan, Henry L Burke, Allen C Smith, Patrick T Kelly, Margaret M Johnson, Rick W Browder, David L Bowton, and Edward F Haponik. Effect on the duration of mechanical ventilation of identifying patients capable of breathing spontaneously. *New England Journal of Medicine*, 335(25):1864–1869, 1996.

Scott K Epstein, Ronald L Ciubotaru, and John B Wong. Effect of failed extubation on the outcome of mechanical ventilation. *Chest*, 112(1):186–192, 1997.

Alexandre Fabregat, Mónica Magret, Josep Anton Ferré, Anton Vernet, Neus Guasch, Alejandro Rodríguez, Josep Gómez, and María Bodí. A machine learning decision-making tool for extubation in intensive care unit patients. *Computer Methods and Programs in Biomedicine*, 200:105869, 2021.

Lucas M Fleuren, Tariq A Dam, Michele Tonutti, Daan P de Bruin, Robbert CA Lalisang, Diederik Gommers, Olaf L Cremer, Rob J Bosman, Sander Rigter, Evert-Jan Wils, et al. Predictors for extubation failure in covid-19 patients using a machine learning approach. *Critical Care*, 25:1–10, 2021.

Fernando Frutos-Vivar, Andrés Esteban, Carlos Apezteguia, Marco González, Yaseen Arabi, Marcos I Restrepo, Federico Gordo, Cristina Santos, Jamal A Alhashemi, Fernando Pérez, et al. Outcome of reintubated patients after scheduled extubation. *Journal of critical care*, 26(5):502–509, 2011.

Joseph Futoma, Morgan Simons, Trishan Panch, Finale Doshi-Velez, and Leo Anthony Celi. The myth of generalisability in clinical research and machine learning in health care. *The Lancet Digital Health*, 2(9):e489–e492, 2020.

Stephanie Godard, Christophe Herry, Paul Westergaard, Nathan Scales, Samuel M Brown, Karen Burns, Sangeeta Mehta, Frank J Jacono, Dalibor Kubelik, Donna E Maziak, et al. Practice variation in spontaneous breathing trial performance and reporting. *Canadian respiratory journal*, 2016 (1):9848942, 2016.

Akshay Goel, Almog Gueta, Omry Gilon, Chang Liu, Sofia Erell, Lan Huong Nguyen, Xiaohong Hao, Bolous Jaber, Shashir Reddy, Rupesh Kartha, et al. Llms accelerate annotation for medical information extraction. In *Machine Learning for Health (ML4H)*, pages 82–100. PMLR, 2023.

Shruti Goradia, Arwa Abu Sardaneh, Sujita W Narayan, Jonathan Penm, and Asad E Patanwala. Vasopressor dose equivalence: a scoping review and suggested formula. *Journal of Critical Care*, 61: 233–240, 2021.

Domenico Luca Grieco, Salvatore Maurizio Maggiore, Oriol Roca, Elena Spinelli, Bhakti K Patel, Arnaud W Thille, Carmen Sílvia V Barbas, Marina Garcia de Acilu, Salvatore Lucio Cutuli, Filippo Bongiovanni, et al. Non-invasive ventilatory support and high-flow nasal oxygen as first-line treatment of acute hypoxemic respiratory failure and ards. *Intensive care medicine*, 47:851–866, 2021.

Margaret S Herridge, Angela M Cheung, Catherine M Tansey, Andrea Matte-Martyn, Natalia Diaz-Granados, Fatma Al-Saidi, Andrew B Cooper, Cameron B Guest, C David Mazer, Sangeeta Mehta, et al. One-year outcomes in survivors of the acute respiratory distress syndrome. *New England Journal of Medicine*, 348(8):683–693, 2003.

Margaret S Herridge, Catherine M Tansey, Andrea Matté, George Tomlinson, Natalia Diaz-Granados, Andrew Cooper, Cameron B Guest, C David Mazer, Sangeeta Mehta, Thomas E Stewart, et al. Functional disability 5 years after acute respiratory distress syndrome. *New England Journal of Medicine*, 364(14):1293–1304, 2011.

Meng-Hsuen Hsieh, Meng-Ju Hsieh, Chin-Ming Chen, Chia-Chang Hsieh, Chien-Ming Chao, and Chih-Cheng Lai. An artificial neural network model for predicting successful extubation in intensive care units. *Journal of clinical medicine*, 7 (9):240, 2018.

Yutaka Igarashi, Kei Ogawa, Kan Nishimura, Shuichiro Osawa, Hayato Ohwada, and Shoji Yokobori. Machine learning for predicting successful extubation in patients receiving mechanical ventilation. *Frontiers in Medicine*, 9:961252, 2022.

Aaron Joffe and Christopher R Barnes. Extubation of the potentially difficult airway in the intensive care unit. *Current Opinion in Anesthesiology*, 35 (2):122–129, 2022.

Alistair EW Johnson, Lucas Bulgarelli, Lu Shen, Alvin Gayles, Ayad Shammout, Steven Horng, Tom J Pollard, Sicheng Hao, Benjamin Moody, Brian Gow, et al. Mimic-iv, a freely accessible electronic health record dataset. *Scientific data*, 10(1): 1, 2023.

Amal Jubran, Brydon JB Grant, Lisa A Duffner, Eileen G Collins, Dorothy M Lanuza, Leslie A Hoffman, and Martin J Tobin. Long-term outcome after prolonged mechanical ventilation. a long-term acute-care hospital study. *American journal of respiratory and critical care medicine*, 199(12):1508–1516, 2019.

Michael Klompas, Deverick Anderson, William Trick, Hilary Babcock, Meeta Prasad Kerlin, Lingling Li, Ronda Sinkowitz-Cochran, E Wesley Ely, John Jernigan, Shelley Magill, et al. The preventability of ventilator-associated events. the cdc prevention epicenters wake up and breathe collaborative. *American journal of respiratory and critical care medicine*, 191(3):292–301, 2015.

James S Krinsley, Praveen K Reddy, and Abid Iqbal. What is the optimal rate of failed extubation? *Critical Care*, 16(1):111, 2012.

Lingyao Li, Jiayan Zhou, Zhenxiang Gao, Wenyue Hua, Lizhou Fan, Huizi Yu, Loni Hagen, Yongfeng Zhang, Themistocles L Assimes, Libby Hemphill, et al. A scoping review of using large language models (llms) to investigate electronic health records (ehrs). *CoRR*, 2024.

Matthew McDermott, Haoran Zhang, Lasse Hansen, Giovanni Angelotti, and Jack Gallifant. A closer look at auroc and auprc under class imbalance. *Advances in Neural Information Processing Systems*, 37:44102–44163, 2024.

Anuj B Mehta, Sohera N Syeda, Renda Soylemez Wiener, and Allan J Walkey. Epidemiological trends in invasive mechanical ventilation in the united states: a population-based study. *Journal of critical care*, 30(6):1217–1221, 2015.

Armand Mekontso-Dessap, Nicolas De Prost, Emmanuelle Girou, François Braconnier, François Lemaire, Christian Brun-Buisson, and Laurent Brochard. B-type natriuretic peptide and weaning from mechanical ventilation. *Intensive care medicine*, 32(10):1529–1536, 2006.

Brenda Y Miao, Christopher YK Williams, Ebenezer Chinedu-Eneh, Travis Zack, Emily Alsentzer, Atul J Butte, and Irene Y Chen. Understanding contraceptive switching rationales from real world clinical notes using large language models. *npj Digital Medicine*, 8(1):221, 2025.

Lorenzo Moja, Koren H Kwag, Theodore Lytras, Lorenzo Bertizzolo, Linn Brandt, Valentina Pecoraro, Giulio Rigon, Alberto Vaona, Francesca Ruggiero, Massimo Mangia, et al. Effectiveness of computerized decision support systems linked to electronic health records: a systematic review and meta-analysis. *American journal of public health*, 104(12):e12–e22, 2014.

Cherubin Mugisha and Incheon Paik. Comparison of neural language modeling pipelines for outcome prediction from unstructured medical text notes. *IEEE Access*, 10:16489–16498, 2022.

Aakanksha Naik, Sravanthi Parasa, Sergey Feldman, Lucy Lu Wang, and Tom Hope. Literature-augmented clinical outcome prediction. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 438–453, 2022.

Stefano Nava, Cesare Gregoretti, Francesco Fanfulla, Enzo Squadrone, Mario Grassi, Annalisa Carlucci, Fabio Beltrame, and Paolo Navalesi. Noninvasive ventilation to prevent respiratory failure after extubation in high-risk patients. *Critical care medicine*, 33(11):2465–2470, 2005.

Takanobu Otaguro, Hidenori Tanaka, Yutaka Igarashi, Takashi Tagami, Tomohiko Masuno, Shoji Yokobori, Hisashi Matsumoto, Hayato Ohwada, and Hiroyuki Yokota. Machine learning for prediction of successful extubation of mechanical ventilated patients in an intensive care unit: a retrospective observational study. *Journal of Nippon Medical School*, 88(5):408–417, 2021.

Lamia Ouanes-Besbes, Fahmi Dachraoui, Islem Ouanes, Rania Bouneb, Faten Jalloul, Mohamed Dlala, Mohamed Fadhel Najjar, and Fekri Abroug. Nt-probnp levels at spontaneous breathing trial help in the prediction of post-extubation respiratory distress. *Intensive care medicine*, 38(5):788–795, 2012.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

Nadeesha Perera, Matthias Dehmer, and Frank Emmert-Streib. Named entity recognition and relation detection for biomedical information extraction. *Frontiers in cell and developmental biology*, 8:673, 2020.

Sebastian Pölsterl. scikit-survival: A library for time-to-event analysis built on top of scikit-learn. *Journal of Machine Learning Research*, 21(212):1–6, 2020.

Hervé Quintard, Erwan l'Her, Julien Pottecher, Frédéric Adnet, Jean-Michel Constantin, Audrey De Jong, Pierre Diemunsch, Rose Fesseau, Anne Freynet, Christophe Girault, et al. Experts' guidelines of intubation and extubation of the icu patient of french society of anaesthesia and intensive care medicine (sfar) and french-speaking intensive care society (srlf) in collaboration with the pediatric association of french-speaking anaesthetists and intensivists (adarpef), french-speaking group of intensive care and paediatric emergencies (gfrup) and intensive care physiotherapy society (skr). *Annals of Intensive Care*, 9:1–7, 2019.

Emily Robitschek, Asal Bastani, Kathryn Horwath, Savyon Sordean, Mark J Pletcher, Jennifer C Lai, Sergio Galletta, Elliott Ash, Jin Ge, and Irene Y

Chen. A large language model-based approach to quantifying the effects of social determinants in liver transplant decisions. *npj Digital Medicine*, 8 (1):665, 2025.

Louise Rose, Michael McGinlay, Reshma Amin, Karen EA Burns, Bronwen Connolly, Nicholas Hart, Philippe Jouvet, Sherri Katz, David Leasa, Cathy Mawdsley, et al. Variation in definition of prolonged mechanical ventilation. *Respiratory care*, 62(10):1324–1332, 2017.

Robert C Rothaar and Scott K Epstein. Extubation failure: magnitude of the problem, impact on outcomes, and prevention. *Current opinion in critical care*, 9(1):59–66, 2003.

Aimee J Sarti, Katina Zheng, Christophe L Herry, Stephanie Sutherland, Nathan B Scales, Irene Watpool, Rebecca Porteous, Michael Hickey, Caitlin Anstee, Anna Fazekas, et al. Feasibility of implementing extubation advisor, a clinical decision support tool to improve extubation decision-making in the icu: a mixed-methods observational study. *BMJ open*, 11(8):e045674, 2021.

Andrew JE Seely, Andrea Bravi, Christophe Herry, Geoffrey Green, André Longtin, Tim Ramsay, Dean Fergusson, Lauralyn McIntyre, Dalibor Kubelik, Donna E Maziak, et al. Do heart and respiratory rate variability improve prediction of extubation outcomes in critically ill patients? *Critical Care*, 18:1–12, 2014.

Di Sun, Lubomir Hadjiiski, John Gormley, Heang-Ping Chan, Elaine Caoili, Richard Cohan, Ajjai Alva, Grace Bruno, Rada Mihalcea, Chuan Zhou, et al. Outcome prediction using multi-modal information: integrating large language model-extracted clinical information and image analysis. *Cancers*, 16(13):2402, 2024.

Arnaud W Thille, Anatole Harrois, Frédérique Schortgen, Christian Brun-Buisson, and Laurent Brochard. Outcomes of extubation failure in medical intensive care unit patients. *Critical care medicine*, 39(12):2612–2618, 2011.

Arnaud W Thille, Jean-Christophe M Richard, and Laurent Brochard. The decision to extubate in the intensive care unit. *American journal of respiratory and critical care medicine*, 187(12):1294–1302, 2013.

Arnaud W Thille, Florence Boissier, Hassen Ben-Ghezala, Keyvan Razazi, Armand Mekontso-Dessap, Christian Brun-Buisson, and Laurent Brochard. Easily identified at-risk patients for extubation failure may benefit from noninvasive ventilation: a prospective before-after study. *Critical Care*, 20(1):48, 2016.

Arnaud W Thille, Grégoire Muller, Arnaud Gacouin, Rémi Coudroy, Maxens Decavèle, Romain Sonneville, François Beloncle, Christophe Girault, Laurence Dangers, Alexandre Lautrette, et al. Effect of postextubation high-flow nasal oxygen with noninvasive ventilation vs high-flow nasal oxygen alone on reintubation among patients at high risk of extubation failure: a randomized clinical trial. *Jama*, 322(15):1465–1475, 2019.

Arnaud W Thille, Florence Boissier, Rémi Coudroy, Sylvain Le Pape, François Arrivé, Laura Marchasson, Jean-Pierre Frat, and Stéphanie Ragot. Sex difference in the risk of extubation failure in icus. *Annals of Intensive Care*, 13(1):130, 2023.

Flavia Torrini, Ségolène Gendreau, Johanna Morel, Guillaume Carteaux, Arnaud W Thille, Massimo Antonelli, and Armand Mekontso Dessap. Prediction of extubation outcome in critically ill patients: a systematic review and meta-analysis. *Critical Care*, 25(1):391, 2021.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

Betty Van Aken, Jens-Michalis Papaioannou, Manuel Mayrdorfer, Klemens Budde, Felix Gers, and Alexander Loeser. Clinical outcome prediction from admission notes using self-supervised knowledge integration. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 881–893, 2021.

Andrew J Vickers and Elena B Elkin. Decision curve analysis: a novel method for evaluating prediction models. *Medical Decision Making*, 26(6):565–574, 2006.

Milena C Vidotto, Luciana CM Sogame, Christiane C Calciolari, Oliver A Nascimento, and

José R Jardim. The prediction of extubation success of postoperative neurosurgical patients using frequency–tidal volume ratios. *Neurocritical care*, 9(1):83–89, 2008.

John E Wennberg. Time to tackle unwarranted variations in practice. *Bmj*, 342, 2011.

Hannah Wunsch, Walter T Linde-Zwirble, Derek C Angus, Mary E Hartman, Eric B Milbrandt, and Jeremy M Kahn. The epidemiology of mechanical ventilation use in the united states. *Critical care medicine*, 38(10):1947–1953, 2010.

Zhixuan Zeng, Xianming Tang, Yang Liu, Zhengkun He, and Xun Gong. Interpretable recurrent neural network models for dynamic prediction of the extubation failure risk in patients with invasive mechanical ventilation in the intensive care unit. *BioData mining*, 15(1):21, 2022.

Qin-Yu Zhao, Huan Wang, Jing-Chao Luo, Ming-Hao Luo, Le-Ping Liu, Shen-Ji Yu, Kai Liu, Yi-Jie Zhang, Peng Sun, Guo-Wei Tu, et al. Development and validation of a machine-learning model for prediction of extubation failure in intensive care units. *Frontiers in medicine*, 8:676343, 2021.

Katina Zheng, Srishti Kumar, Aimee J Sarti, Christophe L Herry, Andrew JE Seely, and Kednapa Thavorn. Economic feasibility of a novel tool to assist extubation decision-making: an early health economic modeling. *International Journal of Technology Assessment in Health Care*, 38(1): e66, 2022.

Sicheng Zhou, Nan Wang, Liwei Wang, Hongfang Liu, and Rui Zhang. Cancerbert: a cancer domain-specific language model for extracting breast cancer phenotypes from electronic health records. *Journal of the American Medical Informatics Association*, 29(7):1208–1216, 2022.

Appendix A. RT Note Distribution

Most encounters had at least one respiratory therapy note collected in the 12 hours prior to extubation. Figure 5 displays number of notes per encounter in the primary cohort of 3,243 encounters.

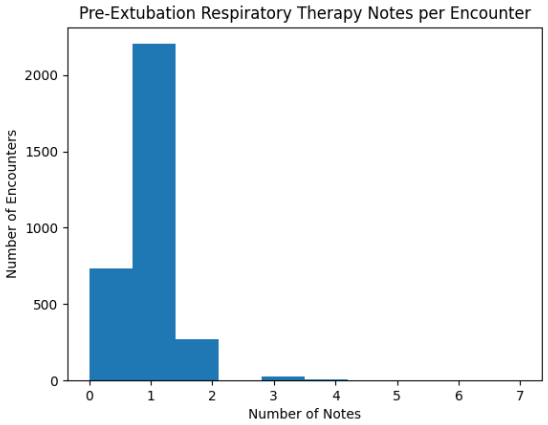


Figure 5: Number of respiratory therapy notes collected within 12 hours prior to extubation

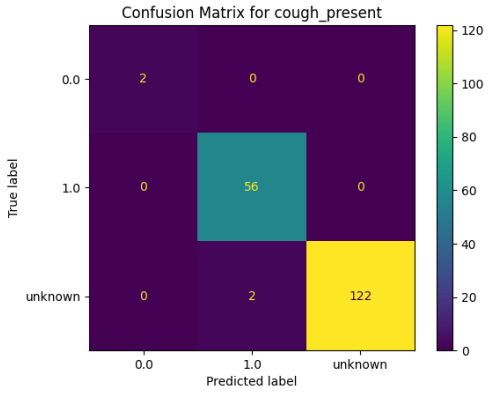
Appendix B. RT Note Extraction Confusion Matrices

Figure 6 contains confusion matrices for the cough features extracted.

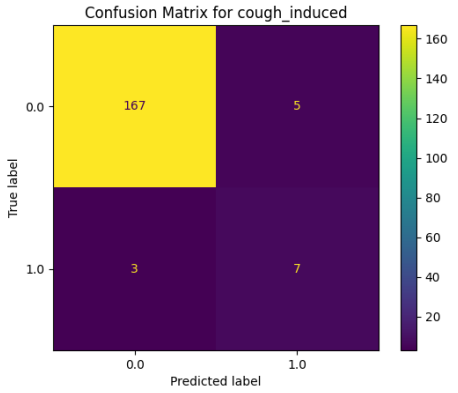
Appendix C. Checklist CDSS Features

The current readiness exam checklist at our institution contains the following predictor variables, of which any two being met indicates that a patient is at high risk for EF:

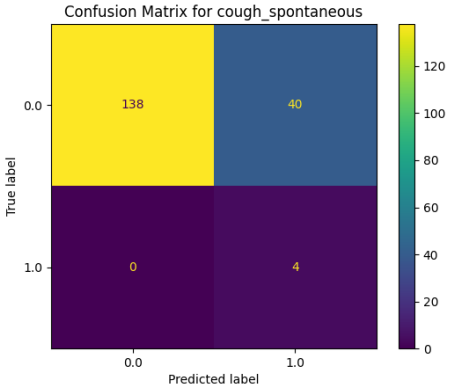
1. History of difficult airway
2. Restricted airway access
3. Concern over reintubation
4. C-spine surgery $>$ 3 levels with operative time $>$ 5h or blood loss $>$ 300 mL
5. Posterior fossa pathology
6. BMI greater than or equal to 40 kg/m^2
7. Lack of cuff leak
8. Lack of spontaneous cough
9. Tracheal suctioning frequency more than twice per hour
10. Frequent oral suctioning
11. Failed more than 3 previous SBTs
12. Age over 60
13. Male gender
14. Coma
15. Chronic lung disease
16. Positive cardiac history
17. End stage kidney disease



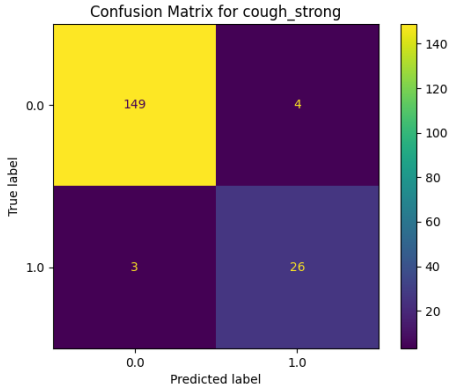
(*a*) Cough Present Confusion Matrix



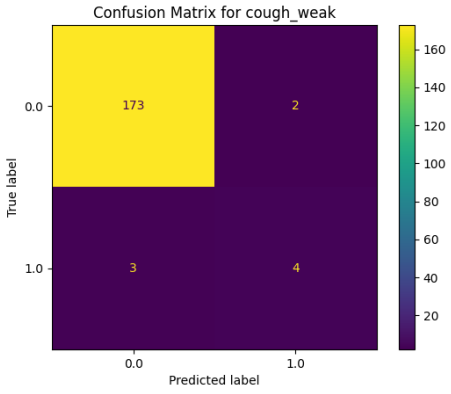
(*b*) Cough Induced Confusion Matrix



(*c*) Cough Spontaneous Confusion Matrix

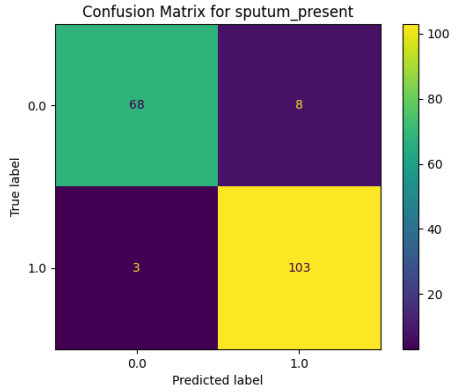


(*d*) Cough Strong Confusion Matrix

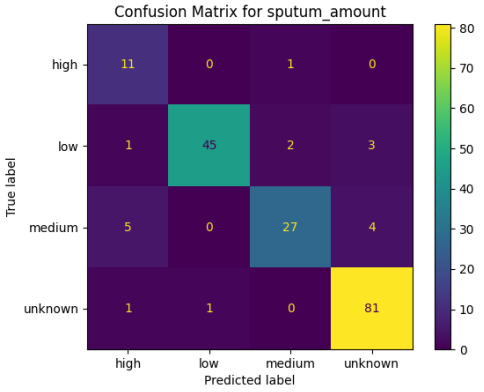


(*e*) Cough Weak Confusion Matrix

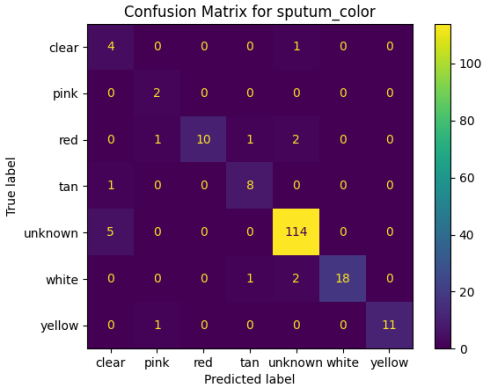
Figure 6: Cough feature confusion matrices for the LLM entity extraction pipeline



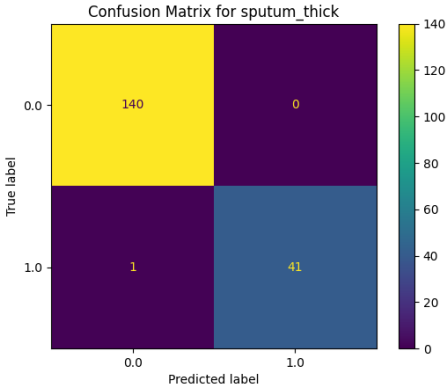
(a) Sputum Present Confusion Matrix



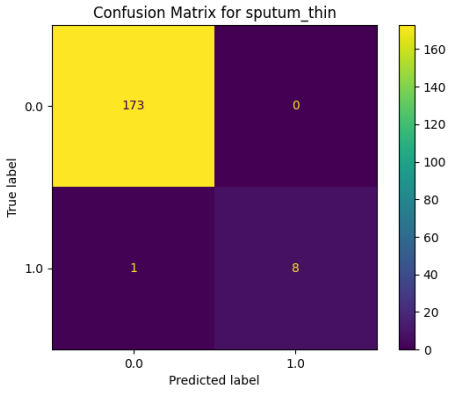
(b) Sputum Quantity Confusion Matrix



(c) Sputum Color Confusion Matrix

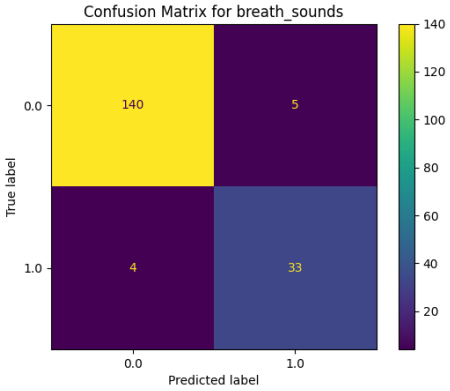


(d) Sputum Thick Confusion Matrix

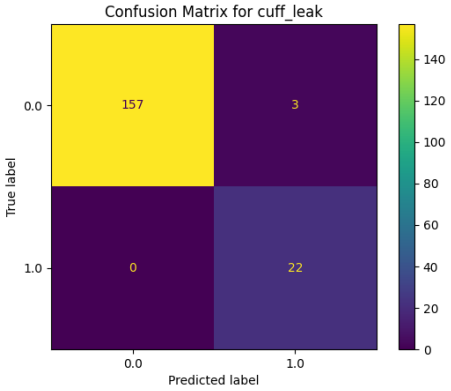


(e) Sputum Thin Confusion Matrix

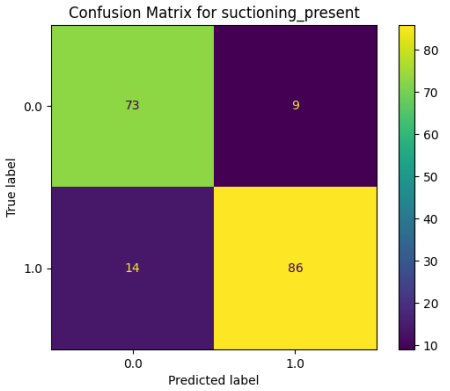
Figure 7: Sputum feature confusion matrices for the LLM entity extraction pipeline



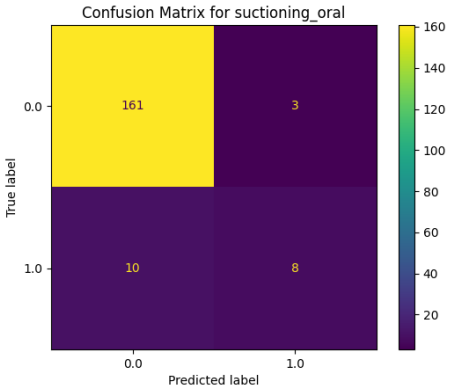
(*a*) Breath Sounds Confusion Matrix



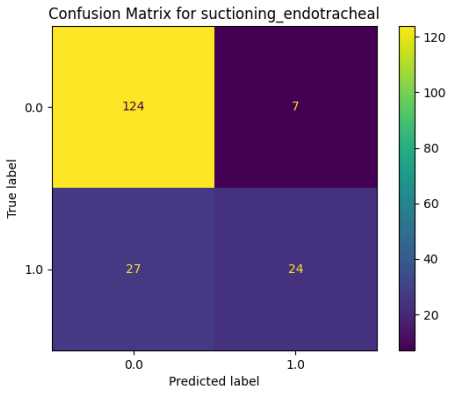
(*b*) Cuff Leak Confusion Matrix



(*c*) Suctioning Present Confusion Matrix



(*d*) Suctioning Oral Confusion Matrix



(*e*) Suctioning Endotracheal Confusion Matrix

Figure 8: Breath Sounds, Cuff Leak, and Suctioning feature confusion matrices for the LLM entity extraction pipeline

		24h Minimum Intubation		1h Minimum Intubation	
Factor	**Group**	N Patients (%)	N Failed (%)	N Patients (%)	N Failed (%)
Gender	Non-Male	1045 (32.22%)	202 (19.33%)	2270 (32.77%)	241 (10.62%)
	Male	2198 (67.78%)	445 (20.25%)	4658 (67.23%)	539 (11.57%)
Age	≤60	2355 (72.62%)	436 (18.51%)	4828 (69.69%)	517 (10.71%)
	>60	888 (27.38%)	211 (23.76%)	2100 (30.31%)	263 (12.52%)
Race/Ethnicity†	Asian	225 (6.94%)	54 (24.00%)	477 (6.89%)	64 (13.42%)
	Black	338 (10.42%)	61 (18.05%)	619 (8.93%)	72 (11.63%)
	Hispanic	309 (9.53%)	51 (16.50%)	606 (8.75%)	65 (10.73%)
	White	2009 (61.95%)	414 (20.61%)	4513 (65.14%)	501 (11.10%)
	Other	362 (11.16%)	67 (18.51%)	713 (10.29%)	78 (10.94%)
Total	-	3243 (100%)	647 (19.95%)	6928 (100%)	780 (11.26%)

Table 5: Patient demographics. Extubation failure rates are computed with a maximum failure window of 7 days. †The Hispanic group includes Hispanic patients of any race and other groups include non-Hispanic patients of any race.

Appendix D. Model Implementation Details

We compute the following medication dosage received by each patient: oral morphine milligram equivalent opioids (Broglio and Portenoy, 2022), diazepam-equivalent benzodiazepine (Ashton, 2002), propofol-equivalent vasopressors (Goradia et al., 2021), crystalloid volume, and a binary variable indicating whether a patient received a neuromuscular blockade.

All input features are normalized to have mean of 0 and unit standard deviation, and missing features are imputed to the training set mean. We measured model performance metrics averaged over 8-fold cross validation of the train set, and we perform a grid search over number of estimators, learning rate, minimum split samples, minimum leaf samples, and tree depth to find optimal hyperparameters for the gradient boosting model, and maximum iterations and L2 regularization strength C for the logistic regression models. We then fit a model using the optimal hyperparameters over all encounters not reserved for testing. We also train and test a model only including RT note features as inputs, which achieves AUROC 0.605 (95% CI: [0.548, 0.658]), indicating predictive value beyond chance.

Appendix E. Predictors for LR Models

The four most important coefficients were common between the LR model trained with a 24 hour minimum IMV duration and a 1 hour minimum duration. These features also are present in prior literature indicating potential predictors for extubation failure. However, several of the less important features are not shared between models, and the sign on the coefficient for FiO_2 measurements differs between models. Table 7 contains a set of the most important coefficients. Sputum amount consistently has the highest magnitude coefficient among the features derived from the RT notes. Table 6 shows summary stats for this cohort.

Below is the full list of variables included in the LR_{B+M+N} model:

- Vitals: mean arterial pressure, heart rate, respiratory rate, SpO_2, and temperature.
- Labs: Anion Gap, Calcium, CO_2 (total), Chlorine, Creatinine, Glucose, Hemoglobin, MCV, Platelet Count, Potassium, Sodium, Urea Nitrogen, White Blood Cell Count, Arterial pH, C-Reactive Protein
- Ventilation Data: Duration of IMV, Number of Spontaneous Breathing Trials, Fio_2, Insp. Flow, Minute Ventilation, Plateau Pressure, Insp. Pressure, AutoPEEP, Observed Tidal Volume

Predictor	Mean (std)	Failed Mean (std)	Non-Failed Mean (std)	P-Value
Initial IMV episode length	93.52 (95.57)	127.41 (117.10)	84.98 (87.30)	$\ll$0.001
Plateau pressure	17.95 (3.72)	18.98 (4.11)	17.69 (3.57)	$\ll$0.001
SpO2	96.90 (1.94)	96.45 (2.12)	97.01 (1.88)	$\ll$0.001
Urea nitrogen	29.12 (22.47)	35.13 (26.80)	27.62 (21.00)	$\ll$0.001
Tempurature	98.60 (1.24)	98.79 (1.27)	98.55 (1.22)	<0.001
Respiration	17.78 (4.16)	18.67 (4.28)	17.56 (4.10)	$\ll$0.001
Number of SBTs	1.39 (1.75)	1.77 (2.02)	1.29 (1.66)	$\ll$0.001
Hemoglobin	9.71 (2.04)	9.34 (1.92)	9.80 (2.06)	<0.001
Sputum amount	0.71 (0.91)	0.85 (0.99)	0.67 (0.89)	0.014
FiO2 (%)	31.33 (9.33)	32.82 (10.12)	30.94 (9.08)	0.003
Predictor	Count (%)	Failed Count (%)	Non-Failed Count (%)	P-Value
Diffuse traumatic brain injury	55 (2%)	22 (4%)	33 (2%)	<0.001
Acute respiratory failure	264 (10%)	32 (6%)	232 (11%)	<0.001
Cerebralvascular disease	344 (13%)	90 (17%)	254 (12%)	0.004
Hemiplegia or paraplegia	221 (9%)	62 (12%)	159 (8%)	0.003
Age > 60	776 (30%)	185 (35%)	591 (28%)	0.003

Table 6: Summary statistics for most important predictors by feature coefficient in LR_{B+M+N} model.

- Diagnoses: Mycocardial Infarction, Congestive Heart Failure, Peripheral Vascular Disease, Cerebrovascular Disease, Dementia, Chronic pulmonary disease, Rheumatic disease, Peptic ulcer disease, Mild liver disease, Diabetes without chronic complication, Diabetes with chronic complication, Hemiplegia or paraplegia, Renal disease, Malignancy, Moderate or severe liver disease, AIDS/HIV, COVID-19, Diffuse traumatic brain injury, Spinal cord injury below neck, Septic shock, Atherosclerotic heart disease, Anemia, Acute respiratory failure, Sleep apnea, COVID-19 exposure, Nicotine dependence (cigarettes), GERD, Hypokalemia, Hypo-osmolality and Hyponatremia, Acute kidney failure, Hyperlipidemia, Hypertension
- Medications: Opioid dose, Benzodiazepine dose, Vasopressor dose, Crystalloid dose, Propofol dose, Neuromuscular Blockade presence
- Demographics: Age over 60, Documented Male Sex
- RT Note Variables: Sputum presence, Sputum thickness, Sputum quantity, Pathological sputum color, Cough presence, Cough strength, Induced cough, Cuff leak, Abnormal breath sounds

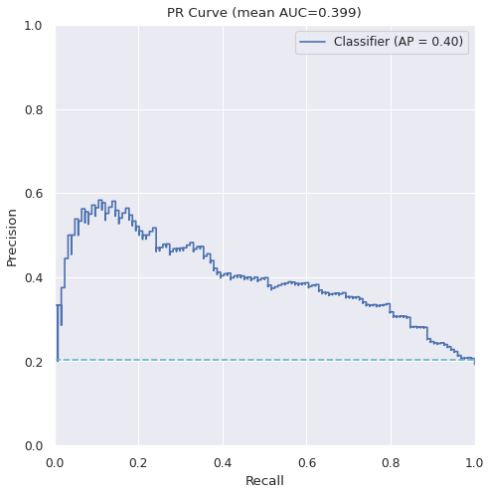


Figure 9: P/R curve for the LR_{B+M+N} model. This model attained an AUROC of 0.75, but fails to achieve precision greater than 0.6 over the test set.

Appendix F. Analysis of Inclusion Criteria

In addition to AUROC, we also measure F1 scores for the positive (patients who did experience EF)

24h Minimum Intubation		**1h Minimum Intubation**	
Predictor	Coefficient	Predictor	Coefficient
Initial IMV Episode Length	0.0955	Initial IMV Episode Length	0.1756
Plateau Pressure (cm H2O)	0.0679	SpO2	-0.1313
SpO2	-0.0664	Urea Nitrogen	0.1051
Urea Nitrogen	0.0601	Plateau Pressure (cm H2O)	0.0815
Diffuse traumatic brain injury	0.0541	Hemoglobin	-0.0721
Temperature	0.0475	Respiratory Rate	0.0720
Respiratory Rate	0.0466	Temperature	0.0687
N SBTs Before Extubation	0.0459	N SBTs Before Extubation	0.0670
Hemoglobin	-0.0434	Sputum amount	0.0620
Acute Resp Failure	-0.0413	**Calcium**	-0.0619
Cerebrovascular disease	0.0410	Cerebrovascular disease	0.0610
Hemiplegia or paraplegia	0.0398	Diffuse traumatic brain injury	0.0588
Sputum amount	0.0392	**Propofol in last 4h**	-0.0522
Age > 60	0.0392	Sputum thickness	0.0475
FiO_2 (%)	0.0372	**Heart rate**	0.0412

Table 7: Variables with the 15 highest magnitude coefficients in the logistic regression BMN model (all features normalized to zero mean and unit standard deviation); note that sputum amount associated with higher risk of extubation failure. Bolded variables are not among the top 20 most important features for the other model. Among all listed features, only FiO_2 has different sign across models.

and negative (patients who did not experience EF) examples in the test set. In the same LR_{B+M+N} model, we observe that negative class F1 tends to decrease and positive class F1 tends to increase as minimum IMV duration increases (Figure 10(a)). However, while a gradient boosting model demonstrates the same trends for AUROC and positive class F1, it does not exhibit the same trend in negative class F1, indicating metric changes may depend on model type (Figure 11). The positive and negative class F1s for the EF Window variation experiments do not vary systematically as EF Window varies (Figure 10(b)).

Appendix G. Demographic Performance Differences

We compute metrics for each demographic subgroup, as shown in Table 8. No performance differences were statistically significant at the $p = 0.05$ significance level, as we lack adequate sample size to demonstrate robust differences in model performance (see Table 5). Both white and non-male patients had lower AUROC than other groups, with lower sensitivity and higher specificity, implying that the LR_{B+M+N} model is less able to identify patients who experienced extubation failure in these groups. While roughly one third of our dataset is non-male, white patients make up more than 60% of our data, meaning performance differences are not related to inadequate representation in training data.

Appendix H. Temporal Generalizability

We assess the intertemporal generalizability of the clinical note variables on an out-of-domain test set. First, we train a logistic regression model on encounters from 2021-2022, including base, medication, and extracted features. Second, we train a model on the same set of encounters, but omit the extracted features. We then assess the fit of both models on a subset of our test split containing encounters from 2023. In this experiment, we determine that performance (based on AUROC) is similar between the model including extracted features (0.751) and that not including such features (0.756). To ensure this

(*a*) Minimum IMV Length Variation; F1

(*b*) EF Window Variation; F1

Figure 10: Changes in F1 for positive and negative class in inclusion variation and EF definition variation experiments. When changing inclusion criteria, positive class F1 increases (primarily due to increases in precision) whereas negative class F1 decreases. Neither class F1 changes greatly as the maximum hours before valid extubation failure changes.

Group	AUROC	AUPRC	Sens. (Recall)	Spec.	PPV (Prec.)	NPV	$F1_+$	$F1_-$	Acc.
All	0.752	0.399	0.734	0.665	0.342	0.913	0.467	0.769	0.678
Non-Male Patients	0.713	0.374	0.676	0.671	0.325	0.898	0.439	0.768	0.672
Male Patients	0.765	0.415	0.759	0.662	0.349	0.920	0.478	0.770	0.681
Age $\leq$ 60	0.749	0.401	0.714	0.681	0.335	0.914	0.456	0.781	0.687
Age $>$ 60	0.754	0.439	0.808	0.581	0.368	0.909	0.506	0.709	0.634
Asian Patients	0.820	0.530	0.931	0.643	0.474	0.964	0.628	0.771	0.717
Black Patients	0.756	0.427	0.633	0.669	0.297	0.892	0.404	0.765	0.663
Hispanic Patients	0.715	0.306	0.750	0.659	0.295	0.933	0.424	0.772	0.673
White Patients	0.673	0.434	0.600	0.675	0.316	0.871	0.414	0.761	0.660
Other Patients	0.748	0.434	0.677	0.676	0.323	0.902	0.438	0.773	0.677

Table 8: Model performance for the LR_{B+M+N} model stratified by mutually exclusive groups. Patients documented as male, who make up 67.78% of our dataset, attain better performance, including better precision and recall, though specificity is improved for patients not documented as male. No differences in metrics are statistically significant at the $p = 0.05$ significance level, as measured over 1000 bootstrap samples of each subset.

(*a*) Minimum IMV Duration Length Variation; AUROC

(*b*) Minimum IMV Duration Length Variation; F1

Figure 11: Experiments in varying initial intubation length, fitting **gradient boosting** models; AUROC and positive class F1 exhibit the same trends as logistic regression models, whereas negative class F1 first decreases, then increases (see Figure 4). This pattern is largely due to changes in specificity: specificity for the model with a 1 hour minimum IMV duration is 0.683, 0.560 for the 12 hour model, and increases to 0.728 for the 17 hour model (these models' respective NPVs are 0.949, 0.934, and 0.896). Increases in positive class F1 are again largely related to increases in precision.

performance was due to the out of domain test set, we also sample a randomly sampled subset of our test split of identical size to the 2021-2022 encounters subset, and measure performance over the 2023 test set for models fit using the extracted features (0.761) and one not using such features (0.754). The 2021-22 subset of our train set includes 1866 examples (72% of our training set); the 2023 test subset includes 193 examples (30% of our test set).

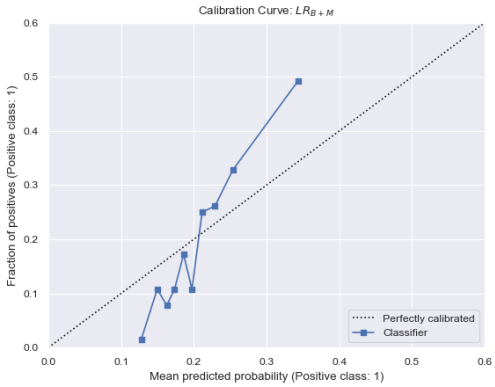
The distributions of outcomes among the 2021-22 patients are quite different from those among the 2023 patients: in the 2021-22 training subset, the extubation failure rate was 21% whereas it was 13% in 2023 (the relatively low EF rate may also explain why the models trained using smaller, out-of-domain sets attained higher AUROC than the models evaluated over all test data). Our finding demonstrates that despite the additional utility of predicting EF in-domain, rapid shifts, such as those between 2021-23 in our institution, may necessitate model retraining, and as Futoma et al. (2020) suggest, the specific use-case of the model should guide its use to predict outcomes for new patients.

Appendix I. Calibration Analysis

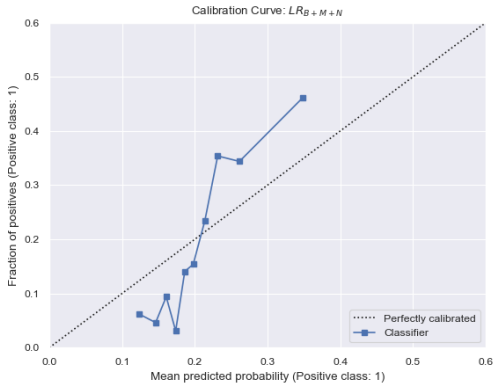
Figure 12 depicts the calibration curves of the LR_{B+M} and LR_{B+M+N} models. Each of these curves has a slope greater than 1, indicating that each model's predictions vary less than the underlying probability of extubation failure. Brier scores were roughly 0.14 for both variants, indicating that inclusion of features from the RT notes does not significantly affect performance.

Appendix J. Decision Curve Analysis

We apply Decision Curve Analysis (Vickers and Elkin, 2006) to assess the net benefit of including additional extracted features when training EF prediction models in terms of additional true positives. This analysis indicates that including extracted features attains increased true positives per false positive at thresholds between 0.17 and 0.22, with net benefit similar at higher or lower thresholds. Figure 13 depicts the decision curves for the LR_{B+M+N} and LR_{B+M} models.



(a) Calibration: LR_{B+M}



(b) Calibration: LR_{B+M+N}

Figure 12: Calibration curves for the LR_{B+M} and LR_{B+M+N} models; both attain similar calibration, with calibration curve slopes > 1 and Brier scores of roughly 0.14.

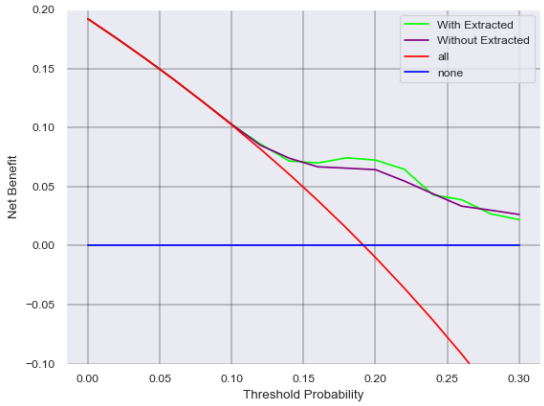


Figure 13: Decision curve analysis comparing the net benefit of the LR_{B+M+N} model to the LR_{B+M} model. The LR_{B+M+N} model ("With Extracted" in the plot), which contains respiratory note features, attains a net benefit of roughly 0.01 true positives higher than the LR_{B+M} model ("Without Extracted" in the plot) at thresholds between 0.17 and 0.22.