Improving ARDS Diagnosis Through Context-Aware Concept Bottleneck Models

Anish Narain ANISH.NARAIN21@IMPERIAL.AC.UK
Imperial College London

Ritam Majumdar R.MAJUMDAR24@IMPERIAL.AC.UK
Imperial College London

Nikita Narayanan NIKITA.NARAYANAN18@IMPERIAL.AC.UK
Imperial College London

Dominic Marshall DOMINIC.MARSHALL12@IMPERIAL.AC.UK
Imperial College London

Sonali Parbhoo S.PARBHOO@IMPERIAL.AC.UK
Imperial College London

Abstract

Large, publicly available clinical datasets have emerged as a novel resource for understanding disease heterogeneity and to explore personalization of therapy. These datasets are derived from data not originally collected for research purposes and, as a result, are often incomplete and lack critical labels. Many AI tools have been developed to retrospectively label these datasets, such as by performing disease classification; however, they often suffer from limited interpretability. Previous work has attempted to explain predictions using Concept Bottleneck Models (CBMs), which learn interpretable concepts that map to higher-level clinical ideas, facilitating human evaluation. However, these models often experience performance limitations when the concepts fail to adequately explain or characterize the task. We use the identification of Acute Respiratory Distress Syndrome (ARDS) as a challenging test case to demonstrate the value of incorporating contextual information from clinical notes to improve CBM performance. Our approach leverages a Large Language Model (LLM) to process clinical notes and generate additional concepts, resulting in a 10% performance gain over existing methods. Additionally, it facilitates the learning of more comprehensive concepts, thereby reducing the risk of information leakage and reliance on spurious shortcuts, thus improving the characterization of ARDS.[1]

1. Introduction

Retrospective identification of Acute Respiratory Distress Syndrome (ARDS) remains a persistent diagnostic challenge in critical care medicine. If cases of ARDS can be accurately identified from large-scale, routinely collected electronic health records (EHR), this would not only yield valuable epidemiological insights but also offer a rich resource for studying disease heterogeneity and treatment response. ARDS is under-recognized, underdocumented and thus poorly coded (Herasevich et al., 2009; Bechel et al., 2023; Poulose et al., 2009). Accurate identification of ARDS is essential in retrospective studies aimed

1. The code is publicly available at https://github.com/ai4ai-lab/context-aware-cbms

at evaluating interventions and outcomes. Automated and retrospective identification of ARDS is challenging as it requires integration of structured tabular data with subjectively interpreted clinical information, such as chest radiographs and unstructured free-text clinical notes (Rubulotta et al., 2024; Ranieri et al., 2012). The current gold standard involves expert case review which is costly and time consuming (Rubulotta et al., 2024). While machine learning (ML) models have shown potential for automating ARDS detection from structured data (Le et al., 2020; Zeiberg et al., 2019), they often fail to capture the full clinical reasoning process and typically offer limited interpretability—posing a major barrier to adoption in healthcare settings.

To address these concerns, Concept Bottleneck Models (CBMs) have been proposed as an interpretable alternative to end-to-end black-box models (Koh et al., 2020). CBMs decompose prediction into two stages: first, the model predicts a set of predefined, human-interpretable concepts from input features; then, these concepts are used to make the final prediction. This architecture enables transparency and the possibility of clinician intervention at the concept level: an important property in safety-critical domains like healthcare. However, despite their promise, CBMs face several limitations that hinder their practical deployment in complex clinical tasks.

A central issue in CBMs is *concept leakage*, where the model learns to infer intermediate concepts using information that is statistically dependent on the target labels, rather than purely from the input features (Mahinpei et al., 2021). This causes the learned concept distribution $p(c|x)$ to become entangled with the label distribution $p(y|x)$, such that concept predictions inadvertently reflect label information. When this happens, the CBM's performance is artificially inflated during training, but the model fails to generalize to out-of-distribution or real-world settings where these dependencies do not hold. In the clinical context, structured concepts such as "$PaO_2/FiO_2 < 300$" may be defined based on, or exhibit strong correlation with, ARDS labels, making them especially susceptible to information leakage.

To mitigate these limitations, we propose a hybrid, context-aware Concept Bottleneck Model that integrates structured EHR data with *LLM-derived concepts* from unstructured clinical notes to improve retrospective diagnosis of ARDS. LLM-derived concepts are generated from sources such as radiology reports, discharge summaries, and echocardiography interpretations, which are authored independently of the labeling process and often reflect rich clinical context. Because LLMs infer these representations directly from descriptive documentation rather than structured variables aligned with labels, the resulting concept distribution $p(c_{\text{text}}|x_{\text{text}})$ is less likely to be conditionally dependent on the outcome y given the input, reducing risk of leakage. This additional information helps the model learn more robust and faithful intermediate representations—ones that are informative, interpretable, and not artificially linked to the target label. Our approach trains a standard CBM on structured clinical variables to capture core physiological features. In parallel, we use a large language model to extract contextual concepts from unstructured notes. These concepts are integrated into the CBM's bottleneck layer, forming a multi-modal representation that combines the structure of traditional CBMs with the contextual richness of clinical text.

Our contributions are as follows: (i) We propose a general framework for augmenting CBMs with context from unstructured data, which can be applied to other clinical use cases

requiring multi-modal reasoning. (ii) We demonstrate improved retrospective identification of ARDS using a real-world ICU dataset, showing gains in predictive performance by 8-10%. (iii) We show that augmenting the concepts using LLM-derived contextual concepts improves the completeness of the concept-space. This reduces the model's reliance on structured variables that may encode spurious correlations with the label, thereby mitigating concept leakage and improving the mutual information between concepts and outcomes. (iv) Finally, our model enables transparent, concept-level reasoning, allowing for interventions on misclassified patients, erroneous concepts or shortcut-induced errors. We demonstrate that targeted corrections of mislabeled or erroneous concepts can recover misclassified cases, further improving concept-label alignment and fostering more reliable model behavior in deployment.

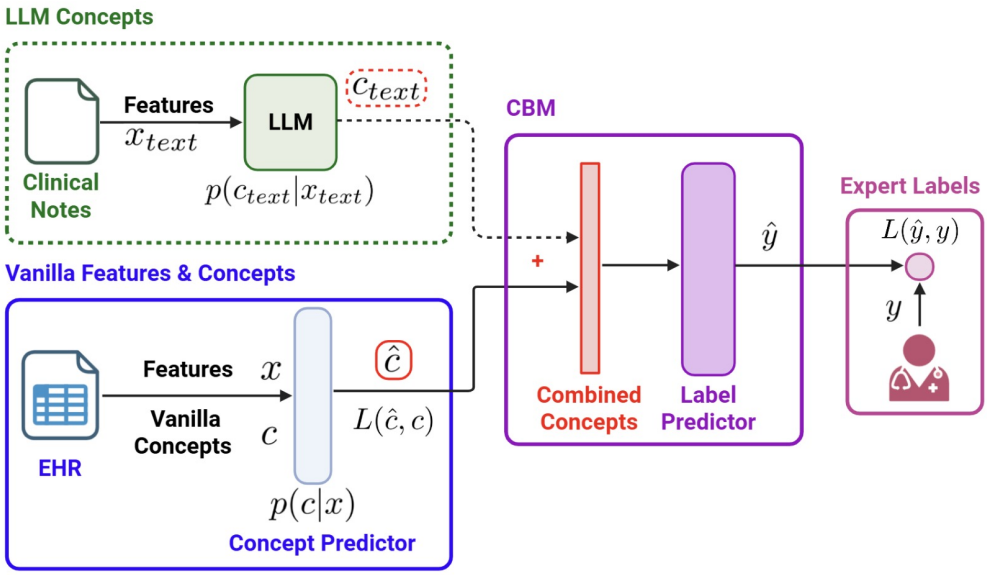

Figure 1: Training pipeline for Context-Aware CBMs. The Vanilla Features & Concepts Box contains the concept predictor, which learns the concept distribution $p(c|x)$ from structured EHR data. We use an LLM (top left) to extract concepts from the unstructured clinical notes, resulting in a separate distribution $p(c_{\text{text}}|x_{\text{text}})$, which differs from $p(c|x)$. These two concept distributions improve the completeness of the concept space and are concatenated to predict the final label y.

Generalizable Insights about Machine Learning in the Context of Healthcare

Our work highlights several broader lessons for the design and deployment of machine learning models in clinical settings, particularly when interpretability and data heterogeneity are key concerns. Although our case study focuses on the retrospective diagnosis of ARDS, it offers broader lessons that are generalizable to a wide range of clinical applications and inform the development of more reliable and actionable ML systems. Specifically,

- *Unstructured Text Offers a Natural Regularizer for Representation Learning.* Text-derived concepts, when generated from physician-authored notes, provide an alternative view of the patient state that is often less correlated with labeling artifacts. These representations, extracted via LLMs, act as a form of distributional regularization — encouraging the model to learn more generalizable and semantically grounded features. That is, LLMs can serve not just as predictors, but as tools for distilling rich, weakly supervised signals that regularize downstream models.
- *Shortcut Learning Can Emerge Even in Human-Defined Features.* Shortcut learning is not limited to raw inputs—our results show that structured, curated clinical variables can also act as shortcuts if they reflect or proxy the label. This extends shortcut risk to features typically seen as safe or interpretable, highlighting the need for diverse, independent data to avoid label-dependent overfitting.
- *Multi-Modal Models Reflect Human-Like Reasoning.* Our model integrates structured and unstructured data to reflect how clinicians combine quantitative metrics with qualitative narratives, improving both predictive robustness and alignment with human reasoning—especially for complex, poorly codified conditions. Multi-modal integration enhances not just performance, but also clinical interpretability.

2. Related Work

Concept Leakage and Shortcut Learning. Several works highlight the problem of concept leakage in CBMs where the models exploit unintended information to improve predictive performance, e.g. (Havasi et al., 2022; Margeloiu et al., 2021). In the same vein, several approaches have been developed to address shortcut learning in deep models (Geirhos et al., 2020; Makar et al., 2022). Some of these methods employ regularization to discourage shortcut variables from influencing predictions (Wang et al., 2022; Makar et al., 2022). But regularization methods often suppress not only spurious features but also causally relevant ones, reducing overall performance (Hong et al., 2025). Others have shown that augmenting a dataset with additional information can help decorrelate shortcuts from data (Cubuk et al., 2019). We build on this idea to show how augmenting concept-based explanations with additional contextual information from unstructured clinical notes using LLMs can reduce a model's reliance on shortcuts and concept leakage overall.

Clinical Concept Bottleneck Models. CBMs are supervised learning models that map raw inputs x to human-understandable concepts c, and then to target labels y, offering insights into the model's reasoning (Margeloiu et al. (2021)). CBMs have been applied across a range of clinical tasks, including predicting vasopressor onset from structured EHR data (Wu et al., 2022), forecasting ICU interventions (Ghassemi et al., 2017), and classifying medical images using interpretable visual markers (Laguna et al., 2024; Yan et al., 2023). Koh et al. (2020) showed that CBMs could identify arthritis by first predicting radiographic features like joint space narrowing. While these models improve interpretability, they often rely on predefined structured features and lack access to the rich contextual information present in clinical narratives. Our work addresses this gap by augmenting CBMs with additional concepts extracted from unstructured clinical notes using LLMs, enabling improved performance on complex retrospective tasks such as ARDS diagnosis.

Clinical Large Language Models, Attention, and Transformers. Attention-based models have been widely employed in healthcare to enhance interpretability by assigning weights to input features, thereby highlighting their contributions to predictive outcomes (De Bois et al., 2021; Chen et al., 2020). However, these models primarily focus on structured data and may not fully capture the rich contextual information present in unstructured clinical notes as we propose in our work. Recent advancements in LLMs have shown promise in clinical settings by extracting insights from unstructured clinical text, such as radiology reports, discharge summaries, and clinical notes. For instance, MedPaLM-2 and Llama-3 have excelled on medical exams (Singhal et al., 2023; Zhou et al., 2023) and BioBERT has been successful in extracting clinical concepts and predicting patient outcomes from free-text data (Lee et al., 2020). Although these models provide powerful insights into unstructured data, they often face limitations such as poor performance in complex cases (Chen et al. (2024)), costly updates, and lack of interpretability, which limits their utility in clinical settings. We address these gaps by using LLMs for concept generation within a CBM, leveraging the interpretability of the models with the text-processing strengths of LLMs. There has also been work to use transformers to predict diseases directly from clinical notes, but this presents key challenges. Standard transformers struggle with long, dispersed documentation due to fixed input lengths (Islam et al. (2025)), leading to truncation-related loss, high computational cost (Vaswani et al. (2017)), and limited ability to capture long-range dependencies (Zhang et al. (2025)). Signals are often diluted by routine content, making raw note processing noisy and inefficient. Our method prompts an LLM to extract predefined, clinically meaningful concepts from unstructured text, yielding interpretable features that retain signal while filtering noise. This reduces complexity for downstream CBMs, improves model focus, and avoids black-box issues in raw-text models.

Methods for ARDS Identification. Current machine learning approaches for identifying ARDS have improved early detection but often lack specificity and generalization. Many models, known as "sniffer systems" (Wayne et al. (2019)), rely on structured EHR data like vital signs and radiology reports but often overfit and fail to generalize across datasets (McKown et al. (2019)). Some use keyword searches in radiology notes (Afshar et al. (2018)), but this lacks adaptability due to institutional variations in terminology. Recent efforts using LLM-based models (Pathak et al. (2023)) show promise in improving generalization by analyzing clinical notes without predefined keywords. Unlike these methods, we leverage the performance of LLMs for text-data and CBMs to deduce a set of interpretable concepts that enable more accurate diagnosis and characterization of ARDS.

3. Cohort

Cohort Selection. We curated a balanced cohort of ARDS and non-ARDS patients from the MIMIC-IV database (Johnson et al., 2023). As with other large critical care datasets, MIMIC-IV lacks an explicit ARDS label. To define the cohort, we selected adult patients ($\geq$ 18 years) with respiratory failure with the following inclusion criteria: PF ratio $<$ 300 with PEEP $\geq$ 5 cm H_2O for three consecutive days, or two days if the patient died on day 3. Exclusion criteria included pregnancy and use of extracorporeal membrane oxygenation (ECMO). These criteria extend prior work, which often used less rigorous methods such as physiological thresholds combined with keyword searches in radiology

reports (Afshar et al. (2018), Gandomi et al. (2022)). Initial validation of those approaches revealed frequent misclassification—especially of patients with transient hypoxia—and failed to exclude cardiac or fluid overload causes of respiratory failure. Our stricter criteria ensure a more reliable enrichment for ARDS before classification. An expert clinician then reviewed clinical records (free-text notes, radiology reports, and echocardiograms) and assigned each patient a label. Of the 1,953 patients reviewed, 1,030 (52.7%) were labeled positive for ARDS, and 923 (47.3%) were labeled negative. Table 1 summarizes cohort statistics.

Variable	**Full Dataset**	**Unseen Data**
n	1953	50
ARDS Positive	1030 (52.7%)	21 (42%)
ARDS Negative	923 (47.3%)	29 (58%)
Female	732 (37.5%)	18 (36%)
Male	1221 (62.5%)	32 (64%)
Avg time in ICU (median)	13.6 days (10.8 days)	13.73 days (10 days)
Age ($\mu \pm \sigma$)	62.6 ± 15.4	65.1 ± 11.6
Ethnicity		
White	1261 (64.6%)	30 (55.6%)
Black or African-American	147 (7.5%)	6 (11.1%)
Hispanic or Latino	75 (3.8%)	1 (1.9%)
Other	131 (6.7%)	6 (11.1%)
Not Available	339 (17.4%)	11 (20.4%)

Table 1: ARDS cohort statistics—comparison between full dataset and unseen data.

Data Preprocessing and Feature Extraction. We curated structured and unstructured EHR data from 1,953 patients in the MIMIC-IV database using subject, stay, and admission identifiers. For structured data, we selected features reflecting both systemic organ dysfunction and respiratory status, in line with clinical characteristics of ARDS. These included components of the SOFA score (respiratory, cardiovascular, renal, central nervous system), vasopressor support (norepinephrine equivalent dose), mechanical ventilation duration, and pre-existing respiratory comorbidities (grouped into six diagnostic categories using ICD-9/10 codes). For each SOFA component, we extracted relevant lab values corresponding to the worst hourly value across the stay, the stay-level average, and the worst value in the first 24 hours. Vasopressor exposure was captured using the time-weighted average and peak norepinephrine dose. Mechanical ventilation duration was calculated across the entire stay. All continuous features were scaled to [0, 1] using min-max normalization, and features with $> 50\%$ missingness were dropped. Remaining missing values were imputed using the median-value imputation, which is more robust to outliers. The full list of features and concepts can be found in Appendix A.

Defining Vanilla Concepts from Tabular EHR Data. We defined a set of vanilla concepts based on the tabular EHR data. These concepts will be used in conjunction

with context-specific concepts derived from free-text data to inform predictions in the next section. The vanilla concepts capture variation in SOFA scores and pre-existing respiratory conditions. SOFA-based concepts included organ-specific SOFA scores as well as the worst hourly SOFA score across a patient's stay, the average SOFA score across the stay and the worst hourly sofa score within the first 24 hours. Concepts based on pre-existing respiratory conditions were derived directly on the basis of condition severity: moderate (1-2 pre-existing conditions) and severe (3+ pre-existing conditions). These concepts are binary. Like the feature set, any missing concept values were imputed using median value imputation and non-binary concepts were standardized to lie between 0 and 1.

Deriving Context-Aware Concepts from Unstructured Clinical Text. Unstructured free clinical text from discharge summaries, radiology reports and echocardiogram studies in MIMIC-IV were used to derive a set of *context-aware* concepts. We use discharge summaries as opposed to admission notes because ARDS typically develops during ICU stay rather than at admission, admission notes often lack diagnostic clarity, and discharge summaries provide a complete view of the clinical trajectory and final diagnoses making them more suitable for retrospective phenotyping. Since the clinical text is authored independently of the labeling process of the patients (their diagnosis), we hypothesize this contextual information may be less susceptible to leakage and reduces the model's reliance on structured variables (shortcuts) that may encode spurious correlations with the label. The concepts we prompt the LLM to extract are designed to capture information that is missing from the structured EHR data. Specifically, from the discharge summaries we extract mentions of physiological events and co-morbidites such as pneumonia, aspiration, pancreatitis, cardiac arrest, and transfusion-related acute lung injury (TRALI). We also ask the LLM to give an overall clinical impression of whether the patient had ARDS based on the textual information alone. In radiology reports, the LLM focused on identifying bilateral infiltrates, while in echocardiogram studies, the LLM identified cases of cardiac failure. LLM prompting details are provided in the next section.

4. Methods

Our goal is to train an interpretable model to classify ARDS by incorporating insights from free-text clinical notes. First, we extract structured EHR data from MIMIC-IV and define features and an initial vanilla concept set that excludes any contextual information derived from clinical notes. We then train an LLM to generate additional concept labels from clinical notes, focusing on previously unused concepts. We then augment the vanilla concept set used with these new LLM concepts into the CBM for context-awareness and use the augmented concept set for downstream prediction. We assess how augmenting the concept set affects prediction accuracy by reducing the CBM's reliance on spurious information. Finally, we discuss how erroneous concepts might be intervened upon at test time and investigate the effect of these interventions on prediction accuracy.

Extracting Context-Aware LLM Concepts From Free-Text. We obtain a set of LLM concepts c_{text} from clinical notes as follows. We implement an LLM, configured with the Llama-3 model (trained on 7 billion parameters), using a chunk size of 4096 tokens and an overlap of 100 tokens between chunks. The prompts we use are based on

earlier investigations but simplified to produce only Yes or No answers, without requesting reasoning behind the responses. After processing the responses, if any contain a "Yes," the concept label is set to 1; otherwise, it remains 0. Notably, the LLM concepts are generated from the distribution $p(c_{\text{text}}|x_{\text{text}})$ and are less likely to be conditionally dependent on the outcome y given the input, thus reducing the risk of leakage.

This work aims to assess whether LLM-generated concepts directly contribute to performance gains. To isolate their effect, we employ a minimal prompt template. Prompt design is known to substantially influence both concept quality and downstream model performance; therefore, using a more complex prompt would confound the source of improvement—whether due to the concepts themselves or the prompt engineering. By fixing the prompt, we control for this source of variability. Moreover, given the high computational cost of generating concepts with LLaMA, a simple prompt also reduces inference time.

We apply the prompt across different patient conditions for clarity and reusability, with minor modifications based on the specific label. For conditions such as aspiration, pneumonia, pancreatitis, cardiac arrest, and TRALI, we use the discharge summary as input and tailor the query to the condition being evaluated. We also use the summaries to get an overall clinical impression of ARDS. For bilateral infiltrates, we modify the query by replacing "suggest" with "mention" and use the radiology reports as input. For cardiac failure, we first assess the echocardiogram studies, and if no failure is diagnosed, we check the discharge summary for any mention of cardiac failure. The prompt format is as follows:

```
template=(
    Context: You are a clinician receiving chunks of clinical
        text for patients in an ICU. Please do the reviewing
       as quickly as possible.
    Task: Determine if the patient had pneumonia.
    Instructions: Answer with 'Yes' or 'No'. If there is not
       enough information, answer 'No'.
    Discharge Text:{discharge_text}
    Query: Does the chunk of text suggest that the patient
       has pneumonia? Answer strictly in 'Yes' or 'No'.
),
input_variables=[discharge_text]
```

The LLM concepts c_{text} are then passed directly as an input to the label predictor.

Problem Setup. Let $(x^{(i)}, y^{(i)}, c_j^{(i)})_{i=1}^{n}$ denote a training set of input features x, target labels y and vanilla concepts c. Here, input $x \in \mathbb{R}^d$ with $d = 21$, consists of 15 continuous features normalized between 0-1 and 6 binary features. Target label $y \in [0, 1]$ are the true classification labels corresponding to ARDS or no ARDS. As our concept set consists of both binary and continuous vanilla concepts, c_j, we define the concepts as follows. For $j \in \{1, 2, \ldots, 12\}$, $c_j \in \mathbb{R}$ as the concepts c_j are continuous. For $j \in \{13, 14\}$, $c_j \in \{0, 1\}$ as the concepts c_j are binary. These concepts are derived from the tabular EHR data as discussed in Section 3. The detailed description of the input features and concepts is present in Appendix A.

Let $f(g(x))$ denote a CBM, where g maps an input x into the concept space and f maps concepts into a final prediction. Let L_Y be the loss function that measures the difference

between the predicted and true target label y values. Let L_{C_j} be a loss function that measures the difference between the predicted and true j-th concept (Koh et al. (2020)). Finally, let the trained CBM $f(g(x))$ be represented using $\hat{f}$ and $\hat{g}$. A joint CBM trains $\hat{f}$ and $\hat{g}$ simultaneously and optimises a weighted sum of the losses for predicting concepts and the target label. That is, for a vanilla CBM:

$$\hat{g}, \hat{f} = \arg\min_{f,g} \sum_i \Big[L_Y(f(g(x^{(i)})); y^{(i)}) + \sum_j \lambda L_{C_j}(g(x^{(i)}); c^{(i)}) \Big] \tag{1}$$

Context-Aware CBM Training. For our Context-Aware CBM, we augment the predicted concepts $\hat{c}$ with the LLM-predicted concepts $c_{\texttt{text}}$ as this gives us additional information that the CBM can use to predict y that is not artificially linked to the label y as it comes from a different distribution. Our initial predicted concepts come from $p(c|x)$ whereas our LLM concepts come from a different distribution $p(c_{\texttt{text}}|x_{\texttt{text}})$. The loss function for the Context-Aware CBM thus becomes:

$$\hat{g}, \hat{f} = \arg\min_{f,g} \sum_i \Big[L_Y(f(g(x^{(i)}), c_{\text{text}}); y^{(i)}) + \sum_j \lambda L_{C_j}(g(x^{(i)}); c^{(i)}) \Big]; \tag{2}$$

Since we incorporate contextual information directly into the label loss of the context-aware CBM, this information serves as a regularization penalty for the concept representation that we learn based on the vanilla concepts. Doing so explicitly constrains the model to prevent it from overfitting to the training data distribution and reduces overreliance on potentially leaky concept information. We use one neural network to predict both intermediate concepts c and final labels y. The network takes two inputs: vanilla features, x, and LLM-predicted concepts, $c_{\texttt{text}}$. The first part of the network g predicts $\hat{c}$ from x (concept predictor in Fig 1). This is concatenated with $c_{\texttt{text}}$ and put through the second part of the network f to predict the label $\hat{y}$ (label predictor in Fig 1). The concept L_{C_j} loss is binary cross entropy (BCE) for binary concepts and mean squared error (MSE) loss for the continuous ones.

5. Interventions in Context-Aware CBMs

CBMs are valued for their interpretability and their support for interventions on misclassified concepts, which can improve downstream label predictions. Koh et al. (2020) show that CBMs enable test-time interventions, allowing users to modify predicted concepts $\hat{c}$, which in turn updates the target label prediction $\hat{y}$. We consider three types of concept-level interventions: (1) Ground truth, where the true concept value is known (ideal but rarely feasible); (2) Mean-based, where the concept is set to its mean; and (3) Median-based, using the concept's median. We primarily explore two strategies for intervening on these concepts.

Interventions assuming concepts are independent. In this scenario, we assume the concept being intervened on does not affect the remaining concepts Koh et al. (2020). This is the most widely used technique, favored for its simplicity and ease of implementation. Mathematically, this is denoted as follows: let $c = [c^1, c^2, ..., c^k, .., c^d]$ be the concept representation before intervention. We are intervening on the k^{th} concept using $c^k_{int} = c^k_{true}/c^k_{mean}/c^k_{median}$. Here, c^k_{true} is the ground-truth value, c^k_{mean} is the mean value

across train examples, while c^k_{median} is the median value across train examples for the concept k. The resulting concept c_{int} after intervention becomes $c_{int} = [c^1, c^2, ..., c^k_{int}, .., c^d]$. This naturally extends across multiple concepts, where a practitioner may choose to intervene on more than 1 concept. In such cases, the concept becomes $c_{int} = [c^1, c^2, ..., c^{k_1}_{int}, ..., c^{k_2}_{int}, .., c^d]$, where k_1 and k_2 are the indices of the concepts being intervened upon.

Interventions assuming correlated concepts. This is more practical, as many concepts are correlated, and assuming independence during intervention can introduce errors: intervening on one concept may require adjustments to others. Vandenhirtz et al. (2024) address this by modeling a joint concept distribution, but this remains a challenging open problem. Here, we propose two techniques based on Pearson correlation between concepts:

1. Correlated interventions based on one main intervention: In this scenario, we have one main concept c^k being actively intervened upon, using ground-truth/mean/median values. Every other concept is intervened as follows: $c^j = c^j + \text{Corr}(c^j, c^k) \times \delta_k$, where $\delta_k = c^k_{int} - c^k$. Here, $\delta_k = c^k_{int} - c^k$ determines the magnitude of the change of concept, while $\text{Corr}(c^j, c^k)$ determines the scale of the change of concept. If $\text{Corr}(c^j, c^k) = 0$, this simplifies to the independent concept intervention.

2. Correlated interventions based on multiple main concept intervention: In this scenario, we have multiple main concepts $c^{k_1}, c^{k_2}, .., c^{k_q}$ being actively intervened upon. The other concepts are intervened as: $c^j = c^j + \text{Corr}(c^j, c^{k_1}) \times \delta_{k_1} + \text{Corr}(c^j, c^{k_2}) \times \delta_{k_2} + \cdots + \text{Corr}(c^j, c^{k_q}) \times \delta_{k_q}$, where $\delta_{k_q} = c^{k_q}_{int} - c^{k_q}$.

These intervention strategies are not exhaustive and can be adapted based on the use case and practitioner expertise. Our experiments illustrate how context-aware CBM performance varies with the choice of intervention.

6. Experimental Setup

1. **Predictive Performance.** We evaluate the predictive performance of both the vanilla CBM and context-aware CBM on our target cohort. These models are compared against standard baseline approaches. In addition, we assess the impact of augmenting baseline models with large language model (LLM)-derived concepts to determine whether incorporating contextual information improves predictive accuracy. 2. **Reducing Leakage.** We investigate whether the context-aware CBM reduces label leakage by mitigating reliance on spurious correlations or shortcut features. To this end, we analyze mutual information and compare it across models. 3. **Interventions Analysis.** We intervene on concepts for incorrectly classified patients using techniques described in Section 5. 4. **Unseen Data.** We test our CBMs on a cohort of data with a different distribution of patients to test the robustness of the models to distribution shift.

7. Evaluation Metrics

We evaluate context-aware CBM performance against vanilla CBMs using accuracy, precision, recall, and F1-score. To assess concept quality, we use mean squared error (MSE) and

mean absolute error (MAE) for continuous concepts, and accuracy and recall for binary concepts. Additionally, we compare mutual information (MI) scores to measure how much information the concepts contribute to the final prediction.

Formally, the MI is defined as:

$$MI(y_{true}; y|c, x) = \sum_{y \in y_{true}} \sum_{\hat{y} \in y|c,x} P(y, \hat{y}) \log \left(\frac{P(y, \hat{y})}{P(y)P(\hat{y})} \right) \tag{3}$$

Here, $P(y, \hat{y})$ is the joint probability of the true label being y and the predicted label being $\hat{y}$, while $P(y)$ and $P(\hat{y})$ are the marginal probabilities of the true and predicted labels, respectively. A MI score of 1 indicates that the concepts capture all relevant information for predicting the final label, whereas a score of 0 suggests that the concepts do not provide any useful information. A lower score also raises the possibility of leakage, where extra input features might bypass the concept layer, undermining the model's interpretability. Mutual information being the evaluation metric for leakage reduction has been theoretically studied in Havasi et al. (2022) and Sun et al. (2024). We point the reader to these references for a more detailed insight into leakage reductions for CBMs. We have also provided additional leakage metrics based on recent work (Parisini et al. (2025)) that considers measuring the leakage in terms of Concept Task Leakage (CTL) and Interconcept leakage (ICL), see Appendix B. We evaluate the effectiveness of the interventions by checking how many false positive and false negative label predictions are corrected following the intervention.

8. Results and Discussion

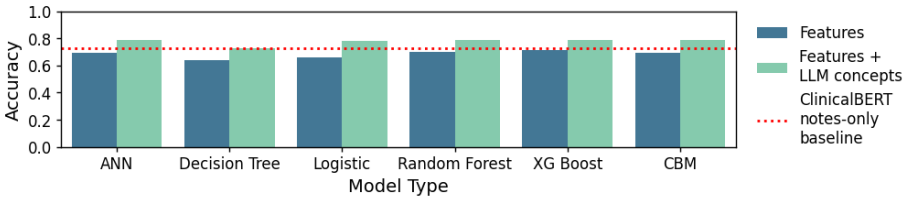

Figure 2: Comparative performance of feature-only models vs. models with LLM-derived concepts. Adding LLM concepts improves all reported metrics by ∼10%. Details on ClinicalBERT baseline in Appendix C. CBMs are competitive with baselines.

Context-aware models augmented with LLM concepts show an 8–10% improvement over traditional models across all metrics. As baselines, we consider two variations of our model architecture: the first uses only physiological signals, while the second incorporates both physiological signals and LLM-extracted features as input. The results are summarized in Fig 2 and Table 10. We observe the inclusion of LLM concepts leads to an 8–10% improvement in ARDS labeling performance across all metrics. However, these baseline models directly map input features to final labels and do not offer insights into the reasoning behind the predictions. Thus, we extend the experiment to CBMs, which serve as the primary model architecture in subsequent experiments. From Table 10, we observe that vanilla CBMs perform comparably to all baselines while offering interpretability.

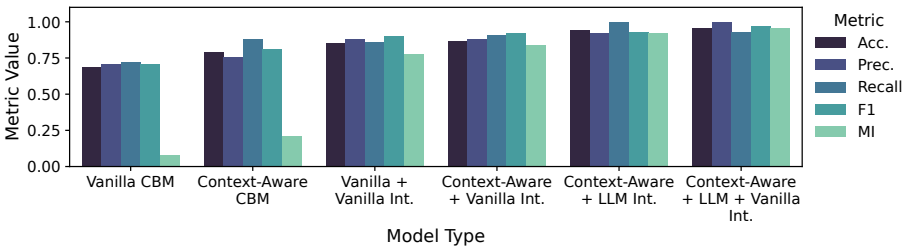

Figure 3: Performance of Vanilla and Context-Aware CBMs after concept-level interventions on misclassified patients. Intervening on the concepts improves metrics by 12–20%, with high gains in mutual information score.

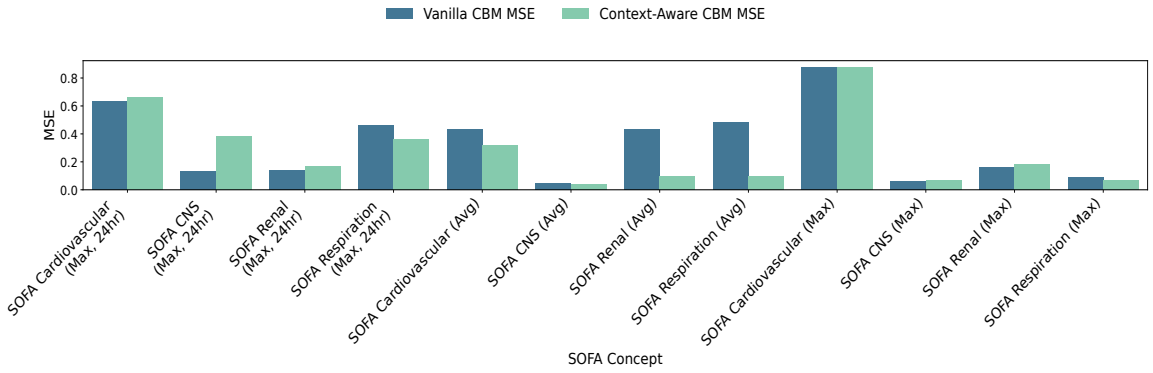

Figure 4: Comparison of concept prediction between Vanilla and Context-Aware CBMs. Context-Aware CBMs achieve lower MSE on respiration-related concepts, indicating better focus on prediction-relevant concepts and reduced leakage.

Context-aware CBMs outperform vanilla CBMs by 10–20% across all metrics, indicating that LLMs extract important concepts relevant for final ARDS prediction. From Fig 3, we observe that context-aware CBMs augmented with LLM-derived concepts improve final ARDS prediction metrics by 10–20%. Furthermore, Table 11 shows gains in predicting SOFA respiration scores and respiratory comorbidities, indicating stronger alignment with clinically relevant features. LLM concepts such as aspiration, bilateral infiltrate, pancreatitis, and the ARDS label are positively correlated with ARDS outcomes, while concepts like cardiac arrests and failures show negative correlations. These additions boost performance over vanilla CBMs. We analyze misclassifications by the vanilla CBM and find that context-aware CBMs correct 78% of false negatives (FN) and 29% of false positives (FP) by leveraging context from clinical notes. However, they also introduce 21% new FP that were previously correct. This reflects a more conservative approach that prioritizes sensitivity, reducing missed ARDS cases at the cost of increased FP. This trade-off is acceptable, as FP can be addressed through follow-up, whereas false negatives risk delayed treatment. Performance on vanilla concepts is also detailed in Table 11.

Table 2: Interventions across multiple concepts. We observe a higher number of corrections when intervening across correlated concepts as compared to intervening independently. Bold indicates better, with higher number of FN/FP corrections per row.

Model	Independent		Correlated	
	FN	FP	FN	FP
Vanilla CBM: GT	24/77	40/46	**38/77**	**42/46**
Vanilla CBM: Mean	26/77	40/46	**26/77**	**42/46**
Vanilla CBM: Median	24/77	40/46	**30/77**	**43/46**
Context-aware CBM: GT	18/22	45/55	**22/22**	**55/55**
Context-aware CBM: Mean	**17/22**	10/55	**17/22**	**31/55**
Context-aware CBM: Median	**17/22**	10/55	**17/22**	**35/55**

Context-aware CBMs improve the prediction of concepts critical to the final outcome, thereby reducing leakage. Figure 4 and Table 11 present statistics on the quality of intermediate concept predictions. We observe that context-aware CBMs achieve lower error on respiratory-related concepts—both continuous (based on SOFA scores) and binary (respiratory comorbidity)—compared to vanilla CBMs. These concepts are crucial for the final ARDS prediction. This suggests that augmenting with LLM-generated concepts helps assign greater emphasis to relevant features, effectively completing the concept space and mitigating information leakage.

Context-aware CBMs produce concepts which have higher mutual information, indicating leakage reduction. From Fig 3, we observe the mutual information of context-aware CBM almost triples, indicating that LLM generated concepts add to the completeness of the concept set and reduce leakage.

Interventions on misclassified concepts further boost the performance of context-aware CBMs by 12–20%. One of the key strengths of CBMs lies in their interpretability through concepts and the ability to intervene on incorrect concept predictions to improve overall model performance (Koh et al., 2020). We evaluate the impact of such interventions in Fig 3, where we observe that intervening on context-aware CBMs results in a performance improvement of 12–20% across all metrics. As a reminder to our readers, we focus on three types of concept-level interventions: a. Ground truth interventions, where the true value of the concept is known. b. Mean-based interventions. c. Median-based interventions. From Tables 2 and Figures 5 & 7, we observe that median-based interventions outperform mean-based ones, correcting more both false positive and false negative cases. This is expected, as the median is more robust to outliers and missing data which are common challenges in medical datasets, whereas the mean is more sensitive to these issues.

Interventions based on correlated concepts are more effective than interventions on individual concepts. In this section, we explore how intervention performance changes when accounting for concept correlations, as opposed to treating concepts independently. While performing multiple interventions, we select the top 3 concepts which had the highest number of corrections from the independent interventions. From Figures 5 & 7, we observe that incorporating correlation awareness leads to substantial improvements, with a significantly higher number of corrections across both false negatives and false positives. Figure F illustrates the correlations among concepts, revealing strong relationships

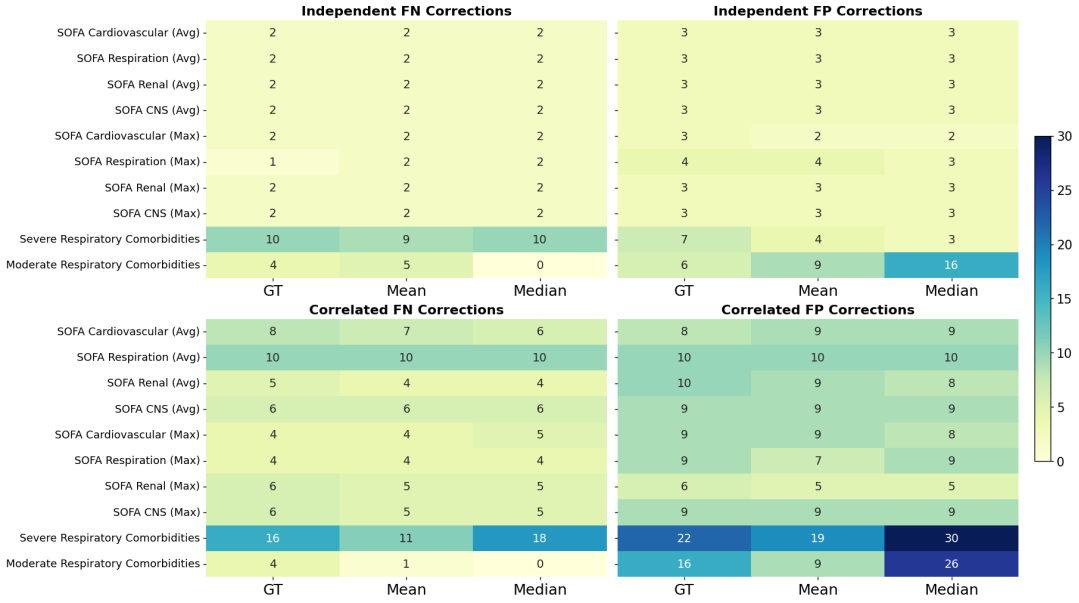

Figure 5: Interventions across individual vanilla concepts. See Figure 7 for LLM concepts. We observe higher number of corrections across correlated interventions as compared to independent interventions. Additionally, median interventions make higher corrections than mean interventions, indicating not all interventions are equally effective and the choice lies on the practitioner.

among various cardiovascular, respiratory, renal, and CNS-related concepts. Furthermore, as shown in Table 2, interventions based on correlated concepts yield greater performance gains than independent interventions for both vanilla and context-aware CBMs.

Context-aware CBMs outperform vanilla CBMs in out-of-distribution patients. We evaluate model performance on an unseen test cohort, as detailed in Table 1. This cohort shows notable demographic shifts, including a 10% decrease in white patients and a 3–4% increase in Black, African-American, Hispanic, Latino, and other racial groups. Additionally, the ARDS class distribution differs from that of the original cohort. As shown in Table 3, context-aware CBMs outperform vanilla CBMs across all metrics, with a 38% improvement in recall. Moreover, their performance remains consistent on the unseen cohort, showing no significant degradation compared to in-distribution test examples. This suggests that context-aware CBMs are robust to distributional shifts.

Discussion on performance boosts. One might assume that adding clinical notes naturally improves performance by simply increasing data volume, but integrating multimodal data is non-trivial (Liu et al., 2022b). Clinical data is heterogeneous: vitals are structured and numerical, whereas notes are unstructured. Simple merging can introduce noise, irrelevant or missing data, and spurious correlations or leakage. We address this by using a LLM

to extract key concepts, grounding inputs in clinically meaningful signals. This improves interpretability, enhances robustness, and enables targeted interventions—benefits often missing from standard multi-modal approaches (Liu et al., 2022a). Aligning structured and unstructured data remains an open problem (Islam et al., 2025), one that requires more than simple concatenation and instead calls for modeling their complementary strengths. A recent ARDS model by Levy et al. (2025) used cTAKES for note processing but observed improvements primarily from radiology reports.

9. Case Studies

Metric	Context-aware	Vanilla
Acc.	**0.80**	0.68
AUC	**0.82**	0.66
Prec.	**0.69**	0.63
Recall	**0.95**	0.57
F1	**0.80**	0.60

Table 3: Performance metrics over unseen test distributions. Context-aware CBMs generalize better to out-of-distribution data than Vanilla CBMs.

We illustrate the interpretability of context-aware CBMs using case studies from our patient cohort. The models expose the key factors driving each diagnosis and let us reason about how changes to specific concepts would alter the predicted diagnosis.

Case study 1: Correction of false negative ARDS classifications by context-aware LLMs. We analyze two patients incorrectly classified as non-ARDS by the vanilla CBM. Patient 1: high vasopressor requirement, low urine output, high average GCS, and no known respiratory diseases or infections. Intermediate concepts show high cardiovascular, renal, respiratory SOFA scores but no severe respiratory comorbidity, leading to a false negative. However, the LLM identifies bilateral infiltrates and pneumonia with no cardiac failure—details absent from physiological signals—correcting the classification to ARDS. Patient 2: average mean BP, high CNS GCS, and lung disease due to external factors. Intermediate concepts indicate high renal and cardiovascular SOFA scores with severe comorbidity. The LLM extracts evidence of aspiration, bilateral infiltrates, and pneumonia with no cardiac failure from clinical notes, enabling correct ARDS classification.

Case study 2: Correction of false positive ARDS classifications by context-aware LLMs. We examine two patients incorrectly classified as ARDS-positive by the vanilla CBM. Patient 1: low vasopressor requirement, moderate urine output, high creatinine, high average GCS, with influenza and chronic lower respiratory disease. Intermediate concepts show high cardiovascular, respiratory, CNS SOFA scores, moderate respiratory comorbidity, and high renal SOFA score in the first 24 hours—leading to a false positive. The LLM identifies cardiac failure, not reflected in physiological signals, correcting the diagnosis to non-ARDS. Patient 2: average mean BP, low urine output, high creatinine, high GCS, with influenza. Predicted concepts include high respiratory, renal, cardiovascular, and CNS SOFA scores with severe comorbidity. The LLM detects cardiac arrest and failure from clinical notes, correcting the classification to negative ARDS.

Case study 3: LLM concepts can induce misclassifications, correctable via concept-level interventions. We analyze two patients where Context-Aware CBMs introduced errors absent in the vanilla CBM predictions. Patient 1: moderate vasopressor requirement and BP, high urine output, low creatinine, high average GCS. Predicted concepts include low cardiovascular and renal SOFA scores, high respiratory SOFA score, and moderate respiratory comorbidity. The vanilla CBM correctly classifies the patient as non-ARDS. However, the LLM detects pneumonia and ARDS in clinical notes, inducing a false positive. Intervening on the ARDS detected concept restores the correct non-ARDS classification. Patient 2: low vasopressor requirement, normal BP, low urine output, high creatinine, with influenza and other respiratory diseases. Predicted concepts include high respiratory, renal, and CNS SOFA scores with severe comorbidity. The vanilla CBM correctly classifies the patient as ARDS-positive. However, the LLM introduces a false cardiac arrest concept, leading to a false negative. Correcting this concept restores the correct ARDS-positive prediction. Hence, we show that even if LLMs introduce leakage, our method makes it easy to fix by leveraging CBMs' ability to trace and edit individual concepts.

Case study 4: Concept-based interpretability enables correction of CBM misclassifications. We examine two patients misclassified by both vanilla and Context-Aware CBMs, highlighting how interpreting intermediate concepts helps identify and correct errors. First, a true positive patient: low vasopressor requirement, normal BP, moderately impaired PaO2-FiO2 ratio, very high urine output, high GCS, with chronic lower respiratory and external agent-induced lung diseases. Predicted correctly: respiratory and CNS SOFA scores; incorrectly: renal SOFA and severe morbidity. LLM identifies aspiration, bilateral infiltrates, and pneumonia. Correcting the morbidity concept aligns both CBMs with the true ARDS-positive label. Second, a true negative patient: low vasopressor requirement, normal BP, moderately impaired PaO2-FiO2 ratio, high urine output, creatinine, and GCS, with other respiratory diseases. Predicted correctly: cardiovascular, respiratory, and renal SOFA scores; incorrectly: CNS SOFA and morbidity. LLM fails to capture cardiac failure. The combined effect of incorrect morbidity and missing cardiac failure leads to misclassification. Intervening on these concepts corrects both CBMs to the true negative label.

10. Conclusions and Limitations

We developed a context-aware concept bottleneck model (CBM) that integrates structured EHR data with unstructured clinical notes processed by a large language model (LLM). This hybrid approach improves both the accuracy and interpretability of retrospective ARDS classification by surfacing clinically meaningful concepts often missed by traditional signals. The CBM's transparency enables physician intervention to correct errors, reducing false positives/negatives, mitigating leakage and shortcuts, improving concept completeness, and strengthening the diagnostic pipeline. Future work includes training models on time-limited data to reflect real-world diagnostic constraints, as our study benefited from full patient timelines. Additionally, while LLM-derived concepts provide valuable signals, they may introduce noise or hallucinations. Hence, clinician oversight will be essential for validation prior to deployment.

References

Majid Afshar, Cara Joyce, Anthony Oakey, Perry Formanek, Philip Yang, Matthew M Churpek, Richard S Cooper, Susan Zelisko, Ron Price, and Dmitriy Dligach. A computable phenotype for acute respiratory distress syndrome using natural language processing and machine learning. In *AMIA Annual Symposium Proceedings*, volume 2018, page 157. American Medical Informatics Association, 2018.

M. A. Bechel, F. Madotto, A. R. Pah, G. Bellani, J. G. Laffey, T. Pham, et al. Validation of a tool for estimating clinician recognition of ards using lung safe data. *PLOS Digital Health*, 2(4):e0000325, 2023. doi: 10.1371/journal.pdig.0000325.

Hanjie Chen, Zhouxiang Fang, Yash Singla, and Mark Dredze. Benchmarking large language models on answering and explaining challenging medical questions. *arXiv preprint arXiv:2402.18060*, 2024.

Peipei Chen, Wei Dong, Jinliang Wang, Xudong Lu, Uzay Kaymak, and Zhengxing Huang. Interpretable clinical prediction via attention-based neural network. *BMC Medical Informatics and Decision Making*, 20:1–9, 2020.

Ekin D Cubuk, Barret Zoph, Dandelion Mane, Vijay Vasudevan, and Quoc V Le. Autoaugment: Learning augmentation strategies from data. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 113–123, 2019.

Maxime De Bois, Mounîm A El Yacoubi, and Mehdi Ammi. Enhancing the interpretability of deep models in healthcare through attention: Application to glucose forecasting for diabetic people. *International Journal of Pattern Recognition and Artificial Intelligence*, 35(12):2160006, 2021.

Amir Gandomi, Phil Wu, Daniel R Clement, Jinyan Xing, Rachel Aviv, Matthew Federbush, Zhiyong Yuan, Yajun Jing, Guangyao Wei, and Negin Hajizadeh. Ardsflag: An nlp/machine learning algorithm to visualize and detect high-probability ards admissions independent of provider recognition and billing codes. *medRxiv*, pages 2022–09, 2022.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020.

Marzyeh Ghassemi, Mike Wu, Michael C Hughes, Peter Szolovits, and Finale Doshi-Velez. Predicting intervention onset in the icu with switching state space models. *AMIA Summits on Translational Science Proceedings*, 2017:82, 2017.

Marton Havasi, Sonali Parbhoo, and Finale Doshi-Velez. Addressing leakage in concept bottleneck models. *Advances in Neural Information Processing Systems*, 35:23386–23397, 2022.

Lena Heidemann, Maureen Monnet, and Karsten Roscher. Concept correlation and its effects on concept-based models. In *2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 4769–4777, 2023. doi: 10.1109/WACV56688.2023.00476.

V. Herasevich, M. Yilmaz, H. Khan, R. D. Hubmayr, and O. Gajic. Validation of an electronic surveillance system for acute lung injury. *Intensive Care Medicine*, 35(6): 1018–1023, 2009. doi: 10.1007/s00134-009-1393-0.

Haoyang Hong, Ioanna Papanikolaou, and Sonali Parbhoo. Do regularization methods for shortcut mitigation work as intended? In *The 28th International Conference on Artificial Intelligence and Statistics*, 2025. URL https://openreview.net/forum?id=TPNFDgltgq.

KM Islam, Ayesha Siddika Nipu, Jiawei Wu, and Praveen Madiraju. Llm-based prompt ensemble for reliable medical entity recognition from ehrs. *arXiv preprint arXiv:2505.08704*, 2025.

Alistair E W Johnson, Tom J Pollard, Lu Shen, Li-wei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. Mimic-iv: a freely accessible electronic health record dataset. *Nature Scientific*, 2023.

Pang Wei Koh, Shiori Sagawa, Hampus Marklund, Sang Michael Xie, Sara Beery Zhang, Akshay Balsubramani, Carla Gomes, Tatsunori Hashimoto, Percy Liang, and Victor Veitch Alvarez. Concept bottleneck models. In *International Conference on Machine Learning (ICML)*, pages 5338–5348. PMLR, 2020.

Sonia Laguna, Ričards Marcinkevičs, Moritz Vandenhirtz, and Julia Vogt. Beyond concept bottleneck models: How to make black boxes intervenable? *Advances in neural information processing systems*, 37:85006–85044, 2024.

Sidney Le, Emily Pellegrini, Abigail Green-Saxena, Charlotte Summers, Jana Hoffman, Jacob Calvert, and Ritankar Das. Supervised machine learning for the early prediction of acute respiratory distress syndrome (ards). *Journal of Critical Care*, 60:96–102, 2020.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240, 2020.

Elizabeth Levy, Dru Claar, Barry D Fuchs, Jennifer Ginestra, Rachel Kohn, Jakob I McSparron, Bhavik Patel, Gary E Weissman, Meeta Prasad Kerlin, Michael W Sjoding, et al. Development and external validation of a detection model to retrospectively identify patients with acute respiratory distress syndrome. *Critical Care Medicine*, 53(6): e1224–e1234, 2025.

Sicen Liu, Xiaolong Wang, Yongshuai Hou, Ge Li, Hui Wang, Hui Xu, Yang Xiang, and Buzhou Tang. Multimodal data matters: Language model pre-training over structured and unstructured electronic health records. *IEEE Journal of Biomedical and Health Informatics*, 27(1):504–514, 2022a.

Ziyi Liu, Jiaqi Zhang, Yongshuai Hou, Xinran Zhang, Ge Li, and Yang Xiang. Machine learning for multimodal electronic health records-based research: Challenges and perspectives. In *China Health Information Processing Conference*, pages 135–155. Springer, 2022b.

Anita Mahinpei, Justin Clark, Isaac Lage, Finale Doshi-Velez, and Weiwei Pan. Promises and pitfalls of black-box concept learning models. *arXiv preprint arXiv:2106.13314*, 2021.

Maggie Makar, Ben Packer, Dan Moldovan, Davis Blalock, Yoni Halpern, and Alexander D'Amour. Causally motivated shortcut removal using auxiliary labels. In *International Conference on Artificial Intelligence and Statistics*, pages 739–766. PMLR, 2022.

Andrei Margeloiu, Matthew Ashman, Umang Bhatt, Yanzhi Chen, Mateja Jamnik, and Adrian Weller. Do concept bottleneck models learn as intended? *arXiv preprint arXiv:2105.04289*, 2021.

Andrew C McKown, Ryan M Brown, Lorraine B Ware, and Jonathan P Wanderer. External validity of electronic sniffers for automated recognition of acute respiratory distress syndrome. *Journal of intensive care medicine*, 34(11-12):946–954, 2019.

Enrico Parisini, Tapabrata Chakraborti, Chris Harbron, Ben D. MacArthur, and Christopher R. S. Banerji. Leakage and interpretability in concept-based models. *arXiv preprint arXiv:2504.14094v2*, May 2025. URL https://arxiv.org/abs/2504.14094v2. 35 pages, 24 figures, revised version v2 dated May 192025.

Ashwin Pathak, Rishi Kamaleswaran, Curtis Marshall, Carolyn Davis, and Philip Yang. Respbert: A multi-site validation of a natural language processing algorithm, of radiology notes to identify acute respiratory distress syndrome (ards). *Authorea Preprints*, 2023.

J. T. Poulose, R. Cartin-Ceba, A. Shoja, C. Trillo-Alvarez, A. Paul, R. Kashyap, et al. Comparison of icd-9 coding with case review for diagnosing ards. *American Journal of Respiratory and Critical Care Medicine*, 179:A4660, 2009. Abstract presented at the ATS International Conference.

V Marco Ranieri, Gordon D Rubenfeld, B Taylor Thompson, Niall D Ferguson, Ellen Caldwell, Eddy Fan, Luigi Camporota, and Arthur S Slutsky. Acute respiratory distress syndrome: the berlin definition. *JAMA: Journal of the American Medical Association*, 307(23), 2012.

Francesca Rubulotta, Sahar Bahrami, Dominic C. Marshall, and Matthieu Komorowski. Machine learning tools for acute respiratory distress syndrome detection and prediction. *Critical Care Medicine*, 52(11):1768–1780, 2024. doi: 10.1097/CCM.0000000000006390.

Karan Singhal, Shekoofeh Azizi, Tu Tu, Shahriar Mahdavi, Jason Wei, Hyung Won Chung, Deepti Jin, and et al. Towards generalist biomedical ai. *Nature*, 622:497–506, 2023.

Ao Sun, Yuanyuan Yuan, Pingchuan Ma, and Shuai Wang. Eliminating information leakage in hard concept bottleneck models with supervised, hierarchical concept learning. *arXiv preprint arXiv:2402.05945*, February 2024. URL https://arxiv.org/abs/2402.05945. Submitted: 3 Feb 2024.

Moritz Vandenhirtz, Sonia Laguna, Ričards Marcinkevičs, and Julia E. Vogt. Stochastic concept bottleneck models, 2024. URL https://arxiv.org/abs/2406.19272.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 30, pages 5998–6008, 2017.

Jiaxuan Wang, Sarah Jabbour, Maggie Makar, Michael Sjoding, and Jenna Wiens. Learning concept credible models for mitigating shortcuts. *Advances in neural information processing systems*, 35:33343–33356, 2022.

Max T Wayne, Thomas S Valley, Colin R Cooke, and Michael W Sjoding. Electronic "sniffer" systems to identify the acute respiratory distress syndrome. *Annals of the American Thoracic Society*, 16(4):488–495, 2019.

Carissa Wu, Sonali Parbhoo, Marton Havasi, and Finale Doshi-Velez. Learning optimal summaries of clinical time-series with concept bottleneck models. In *Machine Learning for Healthcare Conference*, pages 648–672. PMLR, 2022.

An Yan, Yu Wang, Yiwu Zhong, Zexue He, Petros Karypis, Zihan Wang, Chengyu Dong, Amilcare Gentili, Chun-Nan Hsu, Jingbo Shang, et al. Robust and interpretable medical image classifiers via concept bottleneck models. *arXiv preprint arXiv:2310.03182*, 2023.

Daniel Zeiberg, Tejas Prahlad, Brahmajee K Nallamothu, Theodore J Iwashyna, Jenna Wiens, and Michael W Sjoding. Machine learning for patient risk stratification for acute respiratory distress syndrome. *PloS one*, 14(3):e0214465, 2019.

Gongbo Zhang, Zihan Xu, Qiao Jin, Fangyi Chen, Yilu Fang, Yi Liu, Justin F Rousseau, Ziyang Xu, Zhiyong Lu, Chunhua Weng, et al. Leveraging long context in retrieval augmented language models for medical question answering. *npj Digital Medicine*, 8(1): 239, 2025.

Hongjian Zhou, Boyang Gu, Xinyu Zou, Yiru Li, Sam S Chen, Peilin Zhou, Junling Liu, Yining Hua, Chengfeng Mao, Xian Wu, et al. A survey of large language models in medicine: Progress, application, and challenge. *arXiv preprint arXiv:2311.05112*, 2023.

Acknowledgements

We thank Shreya Kamath for designing Figure 1. Your contribution helped define the face of this paper and your fresh perspective was very insightful.

Appendix A. Cohort Selection and Preprocessing

Feature Selection

- **Features from SOFA Scores and Related Lab Measurements:** SOFA scores (Sequential Organ Failure Assessment) are useful in ARDS diagnosis because ARDS is a systemic condition, not just a respiratory issue—it often arises as part of multi-organ dysfunction, typically in critically ill patients. Of the six SOFA components, we selected four considered most relevant for ARDS characterisation: respiratory, cardiovascular, renal, and CNS. The features were extracted from the MIMIC-derived SOFA tables. For each organ component, we included the lab measurements used to calculate the corresponding SOFA score (e.g. creatinine and urine output for renal). Specifically, we extracted the lab values associated with the worst hourly SOFA score across the stay, the average SOFA score across the stay, and the worst hourly score within the first 24 hours. All of these were treated as continuous features.

- **Features from Norepinephrine:** Norepinephrine equivalent dose was included as a feature because it provides a single, standardised measure of the total vasopressor support a patient required. To capture both overall exposure and peak intensity, we calculated the time-weighted average norepinephrine equivalent dose across the stay, as well as the maximum dose administered. These were continuous features.

- **Feature from Mechanical Ventilation Duration:** Another feature included was the total duration the patient spent on mechanical ventilation during their hospital stay, as this serves as a direct indicator of respiratory support needs and is clinically relevant to ARDS severity. This was a continuous feature.

- **Features from Pre-existing Respiratory Conditions:** We also identified whether patients had any pre-existing respiratory conditions by extracting ICD-9 and ICD-10 diagnosis codes from their hospital admissions. These diagnoses were grouped into six respiratory illness categories: upper respiratory infections, influenza and pneumonia, acute lower respiratory infections, chronic lower respiratory diseases, lung diseases due to external agents, and other respiratory diseases. The features were binary.

- **Feature Pre-processing:** We used a mixture of continuous and binary features. We removed 6 features that were missing values for over 50% of patients (5 first 24hr measurements and 1 creatinine measurement, highlighted in Section A). Then, any missing values were imputed using the median calculated from all patients in the training set. We chose to use the median as it is less sensitive to outliers and there are many patients with extreme values on either end of the ranges for several features. All continuous features were then scaled using min-max scaling to be between 0 and 1.

Some features had to be removed as their values were missing for more than 50% of patients.

SOFA Scores - Continuous Values

1. Worst GCS value for CNS
2. Worst first 24hr GCS value for CNS (removed)
3. Average GCS value for CNS
4. Worst creatinine value for Renal (removed)
5. Worst urine output value for Renal
6. Worst first 24hr creatinine value for Renal (removed)
7. Worst first 24hr urine output value for Renal (removed)
8. Average creatinine value for Renal
9. Average urine output value for Renal
10. Worst PaO2/FiO2 ratio for Respiration
11. Worst first 24hr PaO2/FiO2 ratio for Respiration (removed)
12. Average PaO2/FiO2 ratio for Respiration
13. Worst norepinephrine rate for Cardiovascular
14. Worst mean blood pressure for Cardiovascular
15. Worst first 24hr mean blood pressure for Cardiovascular
16. Worst first 24hr norepinephrine rate for Cardiovascular (removed)
17. Average norepinephrine rate for Cardiovascular
18. Average mean blood pressure for Cardiovascular

Norepinephrine Equivalent Dosage - Continuous Values

1. Average Norepinephrine Dosage
2. Peak Norepinephrine Dosage

Mechanical Ventilation Duration - Continuous Value

1. Minutes Patient was on Mechanical Ventilation

Pre-existing Conditions - Binary Values

1. Upper Respiratory Infections – ICD-9: 460–466, ICD-10: J00–J06
2. Influenza and Pneumonia – ICD-9: 480–487, ICD-10: J09–J18

3. Acute Lower Respiratory Infections – ICD-9: 466–469, ICD-10: J20–J22
4. Chronic Lower Respiratory Diseases – ICD-9: 490–496, ICD-10: J40–J47
5. Lung Diseases Due to External Agents – ICD-9: 500–508, ICD-10: J60–J70
6. Other Respiratory Diseases – ICD-9: 510–519, ICD-10: J80–J99

Vanilla Concepts
SOFA - Continuous Values

1. Worst CNS Score across patient stay
2. Worst CNS Score in first 24 hours
3. Average CNS Score
4. Worst Renal Score across patient stay
5. Worst Renal Score in first 24 hours
6. Average Renal Score
7. Worst Respiration Score across patient stay
8. Worst Respiration Score in first 24 hours
9. Average Respiration Score
10. Worst Cardiovascular Score across patient stay
11. Worst Cardiovascular Score in first 24 hours
12. Average Cardiovascular Score

Pre-existing Conditions - Binary Values

1. Moderate pre-existing respiratory comorbidity
2. Severe pre-existing respiratory comorbidity

LLM Concepts
Whether text suggests patient has each of the following conditions (all binary labels):

1. Pneumonia
2. Aspiration
3. Pancreatitis
4. Cardiac Arrest
5. TRALI
6. Overall impression of ARDS
7. Bilateral Infiltrates
8. Cardiac Failure

Appendix B. Additional Leakage Metrics

There is recent work that considers measuring the leakage in terms of Concept Task Leakage (CTL) and Interconcept Leakage (ICL) (Parisini et al., 2025). CTL measures the additional task-relevant information encoded in the predicted concepts that is not present in the ground-truth concepts. ICL measures the extra information each concept prediction carries about other concepts (cross-concept dependence). Empirically, both metrics are sensitive to leakage and strongly correlate with the effectiveness of concept-level interventions. In practice, CTL can boost task performance by allowing the model to exploit task signals embedded in the concept predictors, whereas ICL provides redundant pathways that enable high task scores even when individual concept predictions are poor (Heidemann et al., 2023). Concept-task leakage is defined as follows:

$$\mathrm{CTL}_i(\hat{c}, c, y) = \max\left(0, \frac{I(\hat{c}_i, y)}{H(y)} - \frac{I(c_i, y)}{H(y)}\right)$$

where $\hat{\mathbf{c}}_\mathbf{i}$ is the predicted concepts and $\mathbf{c}_\mathbf{i}$ are the ground truth concepts. This is the difference in mutual information between the predicted and ground truth mutual information (MI) between the concept **i** and the task label. The CTL score is $0 \leq \mathrm{CTL}_i \leq 1$. Note that a negative difference indicates poor concept learning rather than leakage as it means that the learned concepts are less predictive of the task than we would expect from the ground truth.

Interconcept leakage is defined as the pairwise mutual information between concepts **i** and **j**:

$$\mathrm{ICL}_{ij}(\hat{c}, c) = \max\left(0, \frac{I(\hat{c}_i, \hat{c}_j)}{\sqrt{H(\hat{c}_i)\,H(\hat{c}_j)}} - \frac{I(c_i, c_j)}{\sqrt{H(c_i)\,H(c_j)}}\right)$$

This metric again is $0 \leq \mathrm{ICL}_i \leq 1$ and $\mathbf{ICL_{ii}} = \mathbf{0}$ and it quantifies the extra information that each concept encodes about the other concepts. We performed additional experiments to validate our approach in terms of CTL and ICL. CTL did not materially differ between models, indicating similar amounts of task information encoded beyond the ground-truth concepts. In contrast, ICL was lower for the context-aware CBM, indicating reduced cross-concept information sharing. Under the sufficiency rule of Parisini et al. (2025), if either CTL or ICL is higher for model A than model B while the other metric is comparable, then A exhibits greater leakage. Our results imply the context-aware CBM has less leakage than the vanilla CBM. See Table 4 for the CTL and ICL scores.

Appendix C. Clinical Bert Baseline

We also tried extracting the additional concepts from the clinical notes using ClinicalBert instead of the LLaMa model. In order to enable a fair comparison between the two models, we tried to keep the data-preprocessing as similar as possible. For example, we kept the same decision rule: a patient is classified as having concept if any of their chunks is predicted positive. However, some changes had to be made. We set the chunk size to 512 tokens to match BERT's maximum input length. As ClinicalBERT is not a natural language model, it could not be prompted like the LLaMa model had been. We instead computed cosine

Concept	CTL (Vanilla)	CTL (Context-Aware)	ICL (Vanilla)	ICL (Context-Aware)
c_first24hr_sofa_max_cardiovascular	0	0.009	0.0231	0.0135
c_first24hr_sofa_max_cns	0	0.0098	0.0319	0.0083
c_first24hr_sofa_max_renal	0.0235	0.0006	0.0173	0.0147
c_first24hr_sofa_max_respiration	0.0293	0.029	0.0073	0.0200
c_mod_resp_comorbidity	0.0102	0.0152	0	0
c_sofa_avg_cardiovascular	0	0.0427	0.0211	0.0181
c_sofa_avg_cns	0.0063	0.0106	0.0269	0.0045
c_sofa_avg_renal	0.0772	0.0508	0.0112	0.0086
c_sofa_avg_respiration	0	0	0.0083	0.0203
c_sofa_max_cardiovascular	0	0.0089	0.0104	0.0081
c_sofa_max_cns	0.0213	0	0.0235	0.0075
c_sofa_max_renal	0	0.0689	0.0218	0.0094
c_sofa_max_respiration	0	0.0355	0.0111	0.0040
c_svr_resp_comorbidity	0	0	0	0

Table 4: CTL and ICL scores for Vanilla and Context-Aware CBMS across different concepts. 0 is the lowest score possible, and lower scores are better as they indicate less leakage.

similarity scores between each chunk's embedding and two prototype sentences, one which was positive for the condition and one which was negative. For example, 'This patient has pneumonia' and 'This patient does not have pneumonia'. The label was assigned based on which prototype sentence the chunk was more semantically similar to. In Tables 5, 6, and 7, we outline and compare the ClinicalBERT results. The Context-Aware CBM using ClinicalBERT concepts yields only slightly better recall than a Vanilla CBM but performs worse on all other metrics. The Context-Aware CBM using LLaMa-derived concepts outperforms the Context-Aware CBM with BERT concepts on all metrics by 7-24%. This is because ClinicalBERT performs poorly without labels (we do not have ground truth labels for the LLM concepts) and does not perform reasoning like an LLM does, it just produces embeddings from the text.

In Table 7, we show the performance of three ClinicalBERT baselines that were set up as follows. Firstly, for the features only baseline, all features were pre-processed as described in Appendix A. Each feature was transformed into a line like 'Cardiovascular SOFA – worst mean BP 59 mmHg, avg mean BP 61 mmHg'. Each line was tokenized and passed to a pretrained ClinicalBERT model. The model was fine-tuned for binary classification using the ARDS labels in the training dataset. The ClinicalBERT in this setting underperforms compared to all other baselines as it is trained on real, natural language. The sentences generated lack the context of rich clinical notes. For the notes only baseline, each note was split into overlapping chunks. The ClinicalBERT model was fine-tuned for binary classification using the chunked inputs. This ClinicalBERT model achieves competitive accuracy but it is not interpretable as it does not provide any explanations or intermediate concepts. The final baseline uses the ClinicalBERT note embeddings for each patient and concatenates this with the raw tabular features. This was inputted into a simple two layer neural network with a sigmoid activation function for binary classification.

Table 5: Agreement rates between ClinicalBERT and Llama for each LLM concept.

Concept	Agreement Rate (%)
ARDS Mention	50.74
Aspiration	65.95
Bilateral infiltrates	82.33
Cardiac arrest	70.10
Cardiac failure	59.75
Pancreatitis	75.06
Pneumonia	32.67
TRALI	54.66

Concept	Vanilla CBM		Context-Aware CBM	
	MSE	MAE	MSE	MAE
SOFA Cardiovascular (Max, 24hr)	0.63 ± 0.203	0.41 ± 0.081	$\mathbf{0.47 \pm 0.064}$	$\mathbf{0.36 \pm 0.028}$
SOFA CNS (Max, 24hr)	$\mathbf{0.13 \pm 0.032}$	$\mathbf{0.26 \pm 0.037}$	0.16 ± 0.116	0.29 ± 0.129
SOFA Renal (Max, 24hr)	$\mathbf{0.14 \pm 0.065}$	$\mathbf{0.27 \pm 0.062}$	0.51 ± 0.787	0.36 ± 0.182
SOFA Respiration (Max, 24hr)	0.46 ± 0.714	0.38 ± 0.178	$\mathbf{0.10 \pm 0.020}$	$\mathbf{0.24 \pm 0.014}$
SOFA Cardiovascular (Avg)	0.43 ± 0.220	$\mathbf{0.24 \pm 0.059}$	$\mathbf{0.41 \pm 0.323}$	0.30 ± 0.119
SOFA CNS (Avg)	$\mathbf{0.05 \pm 0.058}$	$\mathbf{0.12 \pm 0.089}$	0.28 ± 0.499	0.21 ± 0.167
SOFA Renal (Avg)	0.43 ± 0.689	0.34 ± 0.214	$\mathbf{0.08 \pm 0.022}$	$\mathbf{0.20 \pm 0.014}$
SOFA Respiration (Avg)	0.48 ± 0.865	0.31 ± 0.243	$\mathbf{0.37 \pm 0.493}$	$\mathbf{0.29 \pm 0.262}$
SOFA Cardiovascular (Max)	0.88 ± 0.526	0.37 ± 0.157	$\mathbf{0.62 \pm 0.140}$	$\mathbf{0.28 \pm 0.007}$
SOFA CNS (Max)	$\mathbf{0.06 \pm 0.036}$	$\mathbf{0.14 \pm 0.036}$	0.07 ± 0.060	0.19 ± 0.090
SOFA Renal (Max)	0.16 ± 0.054	0.29 ± 0.084	$\mathbf{0.13 \pm 0.030}$	$\mathbf{0.24 \pm 0.020}$
SOFA Respiration (Max)	$\mathbf{0.09 \pm 0.073}$	$\mathbf{0.23 \pm 0.126}$	0.16 ± 0.232	0.24 ± 0.141
Concept	**Vanilla CBM**		**Context-Aware CBM**	
	Accuracy	**Recall**	**Accuracy**	**Recall**
Respiratory Comorbidity (Moderate)	0.85 ± 0.048	0.96 ± 0.000	$\mathbf{0.89 \pm 0.035}$	$\mathbf{0.95 \pm 0.002}$
Respiratory Comorbidity (Severe)	0.98 ± 0.002	0.95 ± 0.005	$\mathbf{0.98 \pm 0.002}$	$\mathbf{0.95 \pm 0.009}$

Table 6: Concept prediction performance for Vanilla CBM and Context-Aware CBM using ClinicalBERT concepts. The metrics used are MSE and MAE for regression tasks, and accuracy and recall for classification tasks. Bold indicates better performance.

Appendix D. LLM-Derived Concepts Study

We conducted ablation studies to examine how performance changes when using individual LLM concepts in the context-aware CBM. The results in Table 8 show that each concept

Metric	Vanilla CBM	Context-Aware CBM	BERT Features	BERT Notes	BERT Features + Notes
Precision	0.72	0.71	0.53	0.72	**0.75**
Recall	0.69	0.71	**1.00**	0.86	0.88
F1 Score	0.70	0.70	0.69	0.79	**0.81**
Accuracy	0.69	0.68	0.53	0.72	**0.78**

Table 7: Comparison of label prediction performance between Vanilla CBM and Context-Aware CBM using the ClinicalBERT concepts, ClinicalBERT using text generations of the raw features, ClinicalBERT using just the clinical notes, and ClinicalBERT using both clinical note embeddings and raw features concatenated and put through a simple neural network. Bold indicates better performance.

contributes useful signals, with pneumonia and ARDS being especially impactful. Pneumonia yields the highest accuracy, followed by mention of ARDS and bilateral infiltrates. Individually, concepts achieve strong recall (e.g., 0.91 for bilateral infiltrates), but best performance is achieved when they are combined. Removal of even a single concept leads to performance drops, indicating that each captures complementary information.

To validate the LLM-derived concepts, labels for pneumonia, pancreatitis, and cardiac failure were evaluated against expert annotations (full validation was infeasible), see Table 9. The ARDS labels were clinician created and had 86% agreement ($K = 0.76$) on a sample of 100 patients. For pneumonia and cardiac failure, the LLM achieved strong F1 scores (0.71 and 0.67 respectively) and high recall (>0.9), supporting their utility in capturing meaningful clinical signals. Pancreatitis, though much rarer (2.71% prevalence), had high recall (0.96) but low precision (0.24), suggesting that while the LLM may over-include marginal cases, it successfully captures most true positives.

LLM Concept	Accuracy	AUC	Precision	Recall	F1 Score
ARDS Mention	0.73	0.72	0.69	0.89	0.78
Aspiration	0.68	0.67	0.66	0.82	0.73
Bilateral Infiltrates	0.69	0.67	0.65	0.91	0.76
Cardiac Arrest	0.68	0.66	0.65	0.84	0.73
Cardiac Failure	0.70	0.70	0.71	0.75	0.73
Pancreatitis	0.68	0.67	0.65	0.84	0.74
Pneumonia	0.76	0.76	0.75	0.84	0.79
TRALI	0.67	0.65	0.64	0.87	0.74

Table 8: We ran ablation studies to assess the performance impact of individual LLM-derived concepts within the Context-Aware CBM.

LLM Concept	Accuracy	Precision	Recall	F1 Score	Prevalence
Cardiac Failure	0.69	0.53	0.92	0.67	34.38%
Pancreatitis	0.92	0.24	0.96	0.38	2.71%
Pneumonia	0.69	0.56	0.97	0.71	39.80%

Table 9: Three LLM concepts were validated against expert-derived labels.

Appendix E. Additional Case Study

Case study: We consider two patients with the following features. Patient 1: normal blood pressure, moderately impaired PaO2-FiO2 ratio, very high urine output, normal creatinine, no influenza detected, long ventilation duration, and lung disease due to external agents. Predicted intermediate concepts include high cardiovascular, respiratory, renal, and CNS SOFA scores, indicating high morbidity. Patient 2: normal blood pressure, moderately impaired PaO2-FiO2 ratio, normal urine output, very high creatinine, long ventilation duration, and pneumonia-influenza with lung disease due to external agents. Predicted intermediate concepts are similar, with high morbidity. Despite differences in patient features and non-respiratory concepts, the Context-Aware CBM correctly identifies the respiratory symptoms and classifies both patients as having ARDS.

Appendix F. Additional Results

- Table 10 gives the tabulated results that correspond to Figure 2 but with all standard errors included. This experiment was to show that the inclusion of LLM concepts led to improvement on label prediction metrics.
- Table 11 provides the corresponding tabulated results for Figure 4.
- Table 12 compares the performance of joint and sequential CBMs to confirm that leakage does not stem from the training method.

Model Type	Predictor	Accuracy	AUC	Precision	Recall	F1 Score
ANN	Features	0.69 ± 0.002	0.69 ± 0.001	0.71 ± 0.003	0.71 ± 0.010	0.71 ± 0.004
	Features + LLM concepts	$\mathbf{0.79 \pm 0.007}$	$\mathbf{0.79 \pm 0.007}$	$\mathbf{0.78 \pm 0.004}$	$\mathbf{0.84 \pm 0.011}$	$\mathbf{0.81 \pm 0.007}$
Decision Tree	Features	0.64 ± 0.006	0.63 ± 0.006	0.64 ± 0.006	0.64 ± 0.006	0.64 ± 0.006
	Features + LLM concepts	$\mathbf{0.73 \pm 0.004}$	$\mathbf{0.73 \pm 0.004}$	$\mathbf{0.73 \pm 0.004}$	$\mathbf{0.73 \pm 0.004}$	$\mathbf{0.73 \pm 0.004}$
Logistic	Features	0.66 ± 0.011	0.65 ± 0.011	0.65 ± 0.010	0.76 ± 0.020	0.70 ± 0.011
	Features + LLM concepts	$\mathbf{0.78 \pm 0.009}$	$\mathbf{0.78 \pm 0.009}$	$\mathbf{0.74 \pm 0.010}$	$\mathbf{0.90 \pm 0.020}$	$\mathbf{0.82 \pm 0.008}$
Random Forest	Features	0.70 ± 0.007	0.70 ± 0.006	0.70 ± 0.007	0.70 ± 0.007	0.70 ± 0.007
	Features + LLM concepts	$\mathbf{0.79 \pm 0.006}$	$\mathbf{0.78 \pm 0.006}$	$\mathbf{0.79 \pm 0.006}$	$\mathbf{0.79 \pm 0.006}$	$\mathbf{0.78 \pm 0.006}$
XG Boost	Features	0.71 ± 0.007	0.70 ± 0.006	0.71 ± 0.009	0.71 ± 0.007	0.71 ± 0.006
	Features + LLM concepts	$\mathbf{0.79 \pm 0.006}$	$\mathbf{0.79 \pm 0.006}$	$\mathbf{0.79 \pm 0.006}$	$\mathbf{0.79 \pm 0.006}$	$\mathbf{0.79 \pm 0.006}$
Vanilla CBM	Features	0.69 ± 0.011	0.69 ± 0.009	0.72 ± 0.048	0.69 ± 0.120	0.70 ± 0.044
Context-Aware CBM	Features + LLM concepts	$\mathbf{0.79 \pm 0.008}$	$\mathbf{0.78 \pm 0.006}$	$\mathbf{0.76 \pm 0.021}$	$\mathbf{0.89 \pm 0.058}$	$\mathbf{0.82 \pm 0.015}$

Table 10: Corresponding tabulated results for Figure 2. Comparative performance of feature-only models vs. models with LLM-derived concepts. Adding LLM concepts improves all reported metrics by around 10%. Additionally, CBMs perform competitively against baselines while being interpretable and intervenable. Here, bold indicates better.

Concept	Vanilla CBM		Context-Aware CBM	
	MSE	**MAE**	**MSE**	**MAE**
SOFA Cardiovascular (Max, 24hr)	$\mathbf{0.63 \pm 0.203}$	$\mathbf{0.41 \pm 0.081}$	0.66 ± 0.399	0.42 ± 0.141
SOFA CNS (Max, 24hr)	$\mathbf{0.13 \pm 0.032}$	$\mathbf{0.26 \pm 0.037}$	0.38 ± 0.366	0.39 ± 0.150
SOFA Renal (Max, 24hr)	$\mathbf{0.14 \pm 0.065}$	$\mathbf{0.27 \pm 0.062}$	0.17 ± 0.085	0.31 ± 0.098
SOFA Respiration (Max, 24hr)	0.46 ± 0.714	$\mathbf{0.38 \pm 0.178}$	$\mathbf{0.36 \pm 0.307}$	0.40 ± 0.155
SOFA Cardiovascular (Avg)	0.43 ± 0.220	$\mathbf{0.24 \pm 0.059}$	$\mathbf{0.32 \pm 0.117}$	0.27 ± 0.083
SOFA CNS (Avg)	0.05 ± 0.058	$\mathbf{0.12 \pm 0.089}$	$\mathbf{0.04 \pm 0.027}$	0.15 ± 0.059
SOFA Renal (Avg)	0.43 ± 0.689	0.34 ± 0.214	$\mathbf{0.10 \pm 0.033}$	$\mathbf{0.23 \pm 0.069}$
SOFA Respiration (Avg)	0.48 ± 0.865	0.31 ± 0.243	$\mathbf{0.10 \pm 0.160}$	$\mathbf{0.20 \pm 0.195}$
SOFA Cardiovascular (Max)	0.88 ± 0.526	0.37 ± 0.157	0.88 ± 0.435	$\mathbf{0.35 \pm 0.084}$
SOFA CNS (Max)	$\mathbf{0.06 \pm 0.036}$	$\mathbf{0.14 \pm 0.036}$	0.07 ± 0.034	0.19 ± 0.074
SOFA Renal (Max)	$\mathbf{0.16 \pm 0.054}$	0.29 ± 0.084	0.18 ± 0.055	$\mathbf{0.28 \pm 0.060}$
SOFA Respiration (Max)	0.09 ± 0.073	0.23 ± 0.126	$\mathbf{0.07 \pm 0.039}$	$\mathbf{0.18 \pm 0.040}$
Concept	**Vanilla CBM**		**Context-Aware CBM**	
	Accuracy	**Recall**	**Accuracy**	**Recall**
Respiratory Comorbidity (Moderate)	0.85 ± 0.048	0.96 ± 0.000	$\mathbf{0.89 \pm 0.017}$	$\mathbf{0.96 \pm 0.015}$
Respiratory Comorbidity (Severe)	0.98 ± 0.002	0.95 ± 0.005	$\mathbf{0.98 \pm 0.001}$	$\mathbf{0.96 \pm 0.005}$

Table 11: Corresponding tabulated results for Figure 4. Concept prediction performance for Vanilla CBM and Context-Aware CBM using MSE and MAE for regression tasks, and Accuracy and Recall for classification tasks. Better values are highlighted in bold. We observe, Context-aware CBM outperforms Vanilla CBMs across all respiration concepts, which is critical for the final classification of ARDS. This shows, augmenting LLM concepts leads to assigning a higher weight to the relevant concepts, which ultimately improves the final prediction, thereby improving mutual information and reducing leakage. Here, bold indicates better.

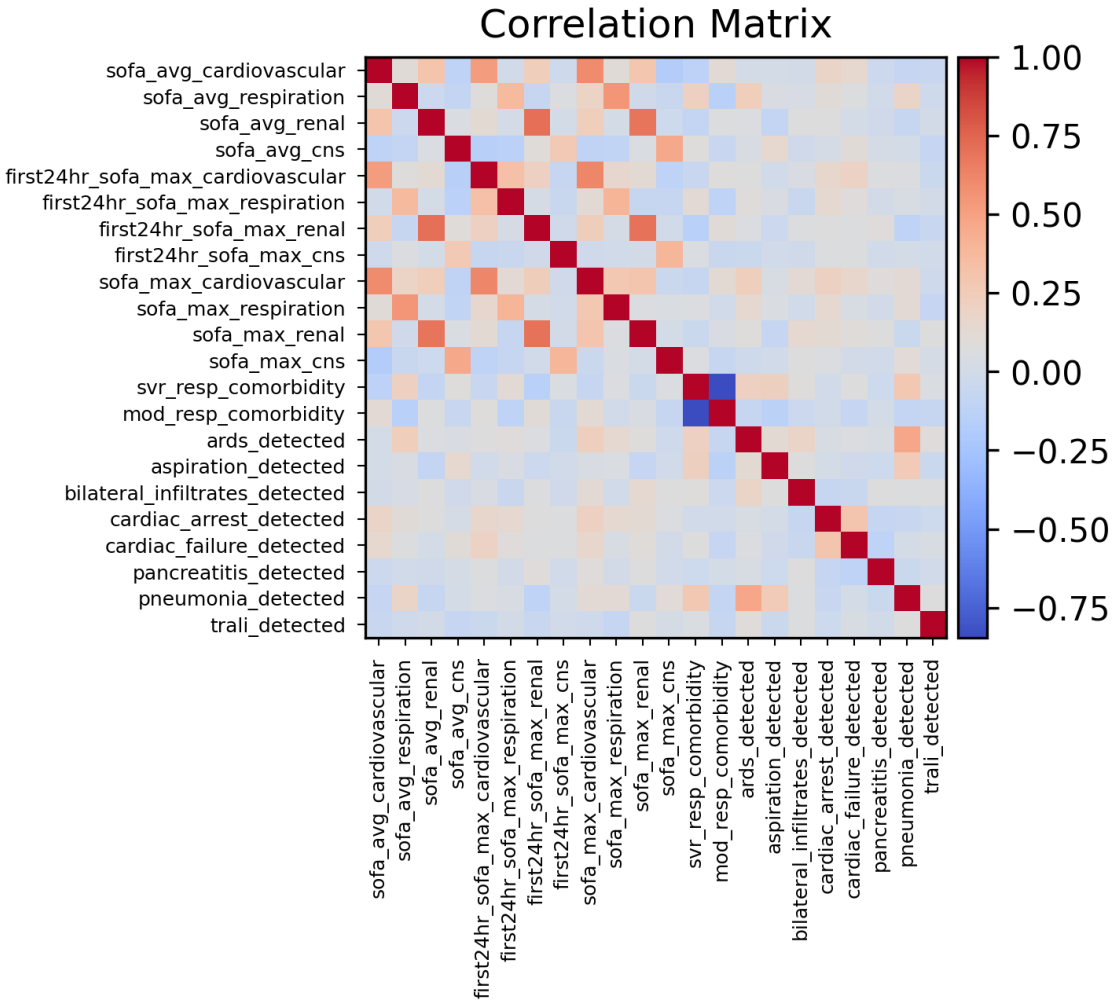

Figure 6: Concept Correlations. We observe strong positive correlations among average cardiovascular SOFA scores, maximum cardiovascular SOFA scores over the first 24 hours, and overall maximum cardiovascular SOFA scores. Similar strong correlations are also observed within the respiration, renal, and CNS-related concepts. Additionally, there is a strong negative correlation between severe and moderate respiratory comorbidities, which aligns with clinical intuition. Among LLM-derived concepts, we find high correlations between impression of ARDS and both pneumonia and bilateral infiltrates, as well as between bilateral infiltrates and pneumonia detection. Because many concepts are strongly correlated, interventions that account for cross-concept dependencies work better than those assuming independence.

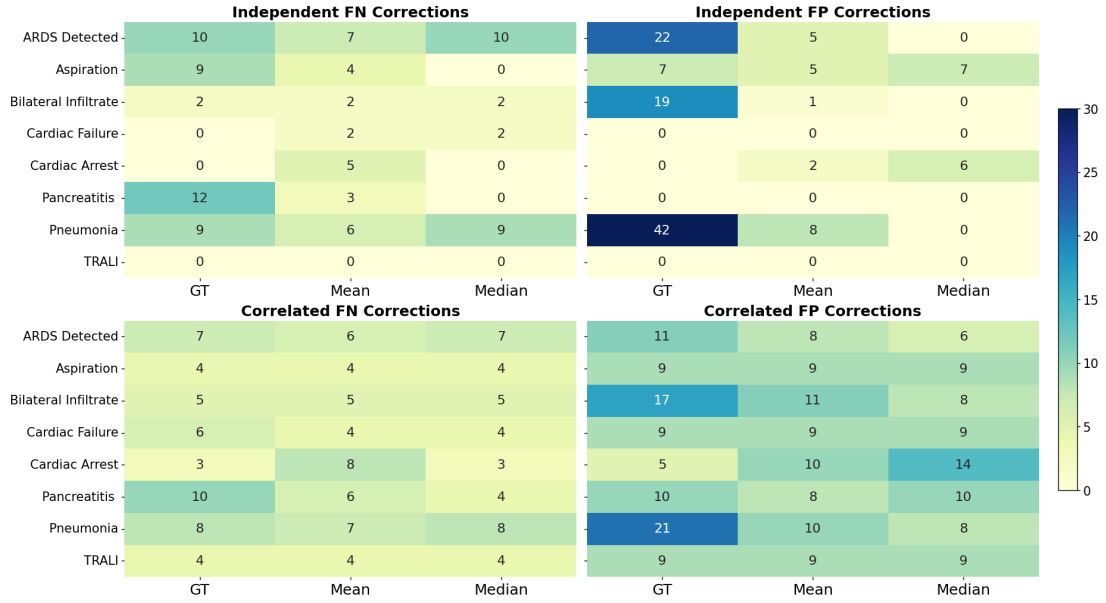

Figure 7: Interventions across individual LLM concepts. See Figure 5 for vanilla concepts. Similar to the vanilla case, we observe more corrections from correlated interventions than from independent ones. However, in this case, the highest number of corrections comes from interventions based on the ground truth method.

LLM Concept	**Acc**	**Prec**	**Rec**	**F1**	**MI**
Vanilla (Joint)	0.69	0.71	0.72	0.71	0.08
Vanilla (Seq)	0.68	0.69	0.74	0.71	0.08
Context-Aware (Joint)	0.79	0.76	0.88	0.81	0.21
Context-Aware (Seq)	0.79	0.77	0.87	0.82	0.21
Context-Aware + Vanilla Intervention (Joint)	0.85	0.88	0.86	0.90	0.78
Context-Aware + Vanilla Intervention (Seq)	0.84	0.82	0.82	0.88	0.76
Context-Aware + LLM Intervention (Joint)	0.94	0.92	1.00	0.93	0.92
Context-Aware + LLM Intervention (Seq)	0.93	0.94	0.98	0.95	0.91
Context-Aware + Vanilla + LLM Inter. (Joint)	0.96	1.00	0.93	0.97	0.96
Context-Aware + Vanilla + LLM Inter. (Seq)	1.00	0.98	0.94	0.98	0.96

Table 12: To test whether label leakage comes from joint training, we compared joint CBMs (concepts and label learned together) with sequential CBMs (learn concepts first, then label). Both strategies have similar performance and MI, suggesting that leakage stems not from joint training but from missing concepts in Vanilla CBMs. Context-Aware CBMs mitigate this by expanding the concept space. Interventions are effective in both settings.