# A Bayesian Counterfactual Early-Warning System for ICU Deterioration Using Irregular Clinical Time Series

Devang Verma
*Independent Researcher*

**Abstract**

Early identification of clinical deterioration in intensive care units (ICUs) remains a persistent challenge due to the irregularity, sparsity, and uncertainty inherent in physiological time series. Delayed recognition of worsening patient states is associated with unplanned ICU transfers, prolonged length of stay, and increased mortality. While traditional early-warning systems based on static thresholds offer interpretability, they are fundamentally limited in their ability to model temporal dynamics, handle informative missingness, or quantify predictive uncertainty in a principled manner. In this work, we propose a missingness-aware, uncertainty-informed early warning framework for modeling time-dependent ICU deterioration risk using irregularly sampled clinical time series from the MIMIC-IV demo dataset. The framework is built upon the GRU-D architecture, which explicitly incorporates observation masks and elapsed-time information into recurrent state updates, enabling robust modeling of clinical trajectories with variable sampling patterns. To quantify epistemic uncertainty, we adopt a Monte Carlo dropout–based variational approximation to Bayesian inference, yielding predictive distributions rather than point estimates of risk. Beyond conventional risk prediction, we introduce a model-implied counterfactual analysis pipeline that simulates hypothetical early interventions through controlled perturbations of physiological trajectories. This allows the framework to examine how predicted risk trajectories evolve under alternative input scenarios, supporting exploratory "what-if" analyses while explicitly avoiding causal treatment effect claims. We further incorporate temporal hold-out evaluation, baseline comparisons, and seed robustness analyses to assess the internal consistency and stability of the proposed approach. Empirical results demonstrate that the model captures clinically plausible temporal patterns of risk escalation and stabilization, exhibits increased predictive uncertainty during physiologically unstable periods, and produces consistent reductions in predicted risk under simulated early intervention scenarios. Cohort-level analyses of peak risk distributions highlight substantial inter-patient heterogeneity, underscoring the potential value of individualized risk trajectories. Although evaluation is limited to a demonstration-scale dataset, the proposed framework establishes a principled methodological foundation for uncertainty-aware and decision-supportive ICU early warning systems, motivating future validation on larger and more diverse clinical cohorts.

# 1 Introduction

Clinical deterioration among hospitalized patients remains a major contributor to preventable morbidity and mortality worldwide. In intensive care units (ICUs), timely recognition of physiological decline is particularly critical, as patient conditions can evolve rapidly and nonlinearly. Delays in escalation of care, even on the order of hours, have been associated with increased rates of organ failure, prolonged hospitalization, and mortality. Consequently, the ability to detect early signs of deterioration before irreversible harm occurs is a central objective of modern critical care practice.[1, 2] To address this need, early-warning systems (EWS) have become a foundational component of hospital safety infrastructure. Conventional EWS, such as the Modified Early Warning Score (MEWS) and the National Early Warning Score (NEWS), rely on manually designed scoring rules applied to a limited set of vital signs. These systems are widely adopted due to their simplicity, transparency, and ease of deployment. However, despite their clinical utility, threshold-based EWS exhibit several fundamental limitations that constrain their effectiveness in complex ICU settings. [3, 4]. First, traditional EWS treat patient observations as isolated snapshots, largely ignoring temporal dependencies and trends that are often clinically informative. Physiological deterioration is rarely abrupt; instead, it manifests as gradual changes, increased variability, or subtle deviations from baseline that evolve over time. Second, these systems depend on fixed thresholds that may not generalize across patient populations, disease subtypes, or institutional practices, leading to reduced sensitivity or excessive false alarms. Third, and critically, conventional EWS provide no principled measure of predictive uncertainty, offering clinicians little insight into the confidence or reliability of generated alerts. Recent advances in machine learning have enabled data-driven approaches to clinical risk prediction using electronic health records (EHRs). Deep learning models, particularly recurrent neural networks, have demonstrated improved performance by explicitly modeling longitudinal patient trajectories and capturing complex temporal dependencies. These approaches have shown promise in ICU deterioration prediction, sepsis detection, and mortality risk estimation [5]. Nevertheless, most existing models remain purely predictive and deterministic in nature. In high-stakes clinical environments, point predictions alone are insufficient; clinicians require systems that communicate uncertainty, support situational awareness, and facilitate informed decision-making under ambiguity.[6]. Another critical limitation of current ICU risk models is the absence of counterfactual reasoning. Predicting that a patient is at high risk does not address the clinically relevant question of whether earlier or alternative interventions could have altered the patient's trajectory. Clinical decision-making is inherently prospective and interventional, involving continuous consideration of "what-if" scenarios related to treatment timing and intensity. Counterfactual analysis provides a structured framework for exploring such hypothetical scenarios by estimating how outcomes might change under alternative conditions. Despite its conceptual importance, counterfactual modeling remains underexplored in ICU early-warning systems, particularly in conjunction with deep temporal models and uncertainty quantification.[7]. In this work, we propose a Bayesian counterfactual early-warning framework for modeling ICU deterioration using irregularly sampled clinical time series. Our approach integrates missingness-aware temporal modeling via the GRU-D architecture with Bayesian inference to quantify epistemic uncertainty in time-dependent risk predictions. By explicitly incorporating observation masks and elapsed-time information, the framework addresses the pervasive irregularity and sparsity of ICU data. We further introduce a model-implied counterfactual simulation mechanism that examines how predicted risk trajectories respond to hypothetical early interventions, enabling exploratory analysis of intervention timing without making causal treatment effect claims. Using the MIMIC-IV demo dataset, we demonstrate that the proposed framework captures dynamic patterns of risk escalation and stabilization, provides uncertainty-aware predictions during physiologically unstable periods, and produces interpretable counterfactual risk reductions under simulated intervention scenarios. While evaluation is necessarily limited in scale, the results highlight the potential of integrating missingness-aware modeling, uncertainty quantification, and counterfactual reasoning within a unified early-warning system. More broadly, this work aims to establish a principled methodological foundation

for decision-supportive ICU risk modeling, motivating future validation on larger and more diverse clinical cohorts.

# 2 Related Work

## 2.1 Traditional Early-Warning Systems

Rule-based early-warning scores such as MEWS and NEWS are widely used in hospital settings due to their simplicity, transparency, and ease of deployment. These systems aggregate a small number of routinely measured vital signs using manually defined thresholds to produce a scalar risk score intended to flag patients at risk of clinical deterioration. Despite their widespread adoption, threshold-based early-warning systems exhibit several well-recognized limitations. Most notably, they treat patient observations as static snapshots, failing to account for temporal trajectories that often precede deterioration. Gradual physiological decline, increasing variability, or subtle deviations from a patient's baseline may therefore go undetected until critical thresholds are crossed. In addition, fixed scoring rules do not adapt to patient-specific baselines or heterogeneous ICU populations, contributing to high false alarm rates and reduced generalizability across institutions and clinical contexts.[4, 3]

## 2.2 Deep Learning for Clinical Time Series

Motivated by the limitations of rule-based systems, recent work has explored data-driven approaches for early warning using electronic health records. Recurrent neural networks, including long short-term memory (LSTM) networks and gated recurrent units (GRUs), have been widely applied to tasks such as mortality prediction, ICU transfer forecasting, and length-of-stay estimation. By explicitly modeling sequential dependencies, these approaches capture temporal patterns that static models based on aggregated features cannot represent.[8] However, standard recurrent architectures implicitly assume regularly sampled inputs and complete observations. In practice, ICU data are highly irregular, with measurement frequency varying across variables, patients, and clinical states. To accommodate this, many deep learning approaches rely on explicit imputation strategies, such as forward filling or population-level mean imputation.[9] While convenient, such strategies may obscure clinically meaningful missingness patterns and conflate unobserved measurements with physiologically normal states, limiting robustness and interpretability in real-world deployments.[5, 10]

## 2.3 Missingness-Aware Temporal Models

To address the challenges posed by irregular sampling and informative missingness, missingness-aware neural architectures have been proposed. Among these, GRU-D extends the standard GRU by introducing decay mechanisms that explicitly model how the influence of past observations diminishes as time elapses. In addition, binary masking vectors encode which variables are observed at each time step, allowing the model to distinguish between measured and unmeasured values. This design enables the network to learn whether the absence of a measurement carries clinical information, rather than treating missingness as noise. Prior studies have shown that incorporating elapsed-time information and observation patterns improves performance on healthcare time-series tasks involving sparse and irregular data. These properties are particularly relevant in ICU settings, where measurement frequency itself often reflects clinical concern and intervention intensity.[9]

### 2.4 Bayesian Deep Learning in Healthcare

Beyond predictive accuracy, uncertainty estimation has emerged as a critical requirement for deploying machine learning models in healthcare. Bayesian deep learning methods aim to quantify predictive uncertainty by approximating distributions over model parameters or outputs, providing information about model confidence alongside point predictions. Such uncertainty estimates are essential in high-stakes environments, where overconfident but incorrect predictions can lead to inappropriate clinical actions.[11] Approximate inference techniques, including Monte Carlo dropout,[6] offer scalable approaches to estimating epistemic uncertainty in deep neural networks. In clinical applications, these methods have primarily been used for static classification tasks, such as mortality prediction at admission or disease detection from single snapshots. Comparatively fewer studies extend Bayesian uncertainty estimation to continuous, time-dependent risk trajectories that evolve as new patient data become available, particularly in conjunction with missingness-aware temporal models.

### 2.5 Counterfactual Reasoning in Clinical AI

Counterfactual inference addresses questions of the form "what would have happened under an alternative scenario," such as earlier or more aggressive intervention. In clinical AI, counterfactual methods are commonly applied to treatment effect estimation, policy evaluation, and retrospective decision analysis. These approaches provide a pathway toward moving beyond risk prediction and supporting actionable clinical reasoning.[7, 12] However, most counterfactual frameworks rely on static covariates, simplified treatment assignments, or strong causal assumptions that are difficult to satisfy in dynamic ICU environments. Integrating counterfactual reasoning with deep temporal models remains an open challenge, particularly in the presence of time-varying confounding, irregular observations, and evolving patient states. Addressing this gap is necessary for developing early-warning systems that not only identify risk but also support reasoning about intervention timing and potential preventability.

### 2.6 Positioning of the Present Work

The present study builds upon these lines of research by integrating missingness-aware temporal modeling, Bayesian uncertainty estimation, and model-implied counterfactual analysis within a unified early-warning framework. Unlike traditional early-warning scores, the proposed approach models continuous risk trajectories over time. In contrast to prior deep learning–based ICU models, it explicitly accounts for irregular sampling and quantifies epistemic uncertainty. Finally, rather than limiting analysis to correlational predictions, the framework explores counterfactual risk trajectories to support exploratory decision-oriented analysis, while explicitly avoiding causal treatment effect claims.

## 3 Dataset and Preprocessing

### 3.1 Dataset Description

This study uses the publicly available MIMIC-IV demo dataset, a de-identified subset of the Medical Information Mart for Intensive Care (MIMIC-IV) database comprising longitudinal electronic health records from intensive care unit (ICU) admissions. The dataset includes demographic attributes, vital signs, laboratory measurements, and clinical outcomes recorded throughout hospital stays. Although limited in cohort size, the demo version preserves the structural characteristics of the full dataset, including irregular sampling patterns, heterogeneous variable availability, and clinically meaningful missingness. The objective of this work is methodological validation rather than benchmark performance. Accordingly, the MIMIC-IV demo dataset is used to evaluate the internal consistency, stability, and qualitative behavior of the proposed

modeling framework under realistic data conditions. Each ICU stay is treated as an independent temporal sequence, and no patient appears in more than one data split.

### 3.2 Temporal Framing and Discretization

Clinical measurements in ICU settings are inherently irregular and asynchronously recorded across variables. To enable temporal modeling while preserving temporal resolution, patient trajectories are discretized into fixed-width hourly time bins beginning at the time of ICU admission. All observations recorded within a given hour are aggregated using clinically appropriate summary statistics, such as the mean for continuous variables. Crucially, if no measurement is recorded for a variable within a given time bin, the value is explicitly treated as missing rather than imputed at this stage. This design choice preserves the sparsity structure of the data and enables downstream models to reason explicitly about observation patterns and elapsed time since last measurement.

### 3.3 Representation of Observations, Missingness, and Time Gaps

For each patient trajectory and each clinical variable $d$ at discrete time step $t$, a triplet representation is constructed:

$$(x_{t,d},\, m_{t,d},\, \Delta t_{t,d}),$$

where $x_{t,d}$ denotes the observed value of variable $d$ at time $t$ if available, $m_{t,d} \in \{0,1\}$ is a binary indicator specifying whether the variable was observed at that time step, and $\Delta t_{t,d}$ represents the elapsed time since the most recent prior observation of variable $d$. The elapsed-time variable $\Delta t_{t,d}$ is defined recursively as the difference between the current time step and the most recent time step $t' < t$ for which the variable was observed:

$$\Delta t_{t,d} = t - \max\left\{t' < t \;:\; m_{t',d} = 1\right\}.$$

This representation explicitly distinguishes unobserved measurements from physiologically normal values and enables downstream models to reason about both observation patterns and temporal gaps between measurements. Such distinctions are critical in ICU data, where the absence of a measurement often reflects clinical judgment rather than random missingness.

### 3.4 Handling of Missing Values and Numerical Stability

Missing values in the raw observation tensor are not directly passed to the predictive model. Instead, missing entries are initially replaced with feature-wise population means computed across the training data solely for numerical compatibility. The associated mask and elapsed-time tensors ensure that the model is informed of which values were imputed and how stale the underlying measurements are. All continuous variables are normalized using z-score normalization based on statistics computed from the training split only. These normalization parameters are fixed and reused for validation and testing to prevent information leakage. Elapsed-time values are clipped to a fixed maximum to prevent numerical instability during downstream exponential decay computations.

### 3.5 Label Construction and Temporal Dependency

The prediction task is formulated as time-dependent risk estimation rather than static classification. At each time step, a binary label indicates whether an ICU deterioration event—operationalized as ICU transfer or a clinically defined proxy—occurs within a predefined future horizon. This formulation produces temporally correlated labels within each patient trajectory and reflects the prospective nature of early warning systems. Patients who do not experience the event within the observation window are treated as right-censored. As

labels are defined at multiple time steps per patient, evaluation metrics are interpreted with caution and are primarily used to assess internal consistency rather than absolute predictive performance.

### 3.6 Data Splitting and Evaluation Integrity

To prevent patient-level leakage, all data splits are performed at the patient level prior to model development. Training, validation, and test sets are constructed such that no individual patient appears in more than one split. These splits remain fixed across all experiments, including baseline comparisons and robustness analyses. In addition to standard patient-level splitting, a temporal hold-out protocol is employed in selected experiments. Models are trained on early portions of each patient trajectory and evaluated on future time windows, providing a more conservative assessment of temporal generalization and mitigating leakage arising from temporally correlated labels.

## 4 Methodology

### 4.1 Problem Formulation

We consider the task of time-dependent risk estimation for clinical deterioration in intensive care unit (ICU) patients. Each patient encounter is represented as a multivariate clinical time series observed over discrete time steps indexed by

$$t = 1, 2, \ldots, T.$$

At each time step, the patient state is described by a vector of physiological variables. We denote the observed trajectory as

$$X_{1:T} = \{x_1, x_2, \ldots, x_T\},$$

where each observation

$$x_t \in \mathbb{R}^D$$

corresponds to measurements of $D$ D clinical variables at time $t$. The modeling objective is to estimate a time-varying risk trajectory, defined as the conditional probability of ICU deterioration occurring within a predefined future horizon:

$$P(Y_t = 1 \mid X_{1:t}),$$

where

$$Y_t \in \{0, 1\}$$

This formulation departs from static classification by producing a continuous sequence of probabilistic risk estimates, enabling early warning, longitudinal monitoring, and retrospective assessment of evolving patient states.

### 4.2 Representation of Missingness and Temporal Irregularity

Clinical time series in ICU environments are characterized by irregular sampling, asynchronous measurements across variables, and informative missingness. To explicitly model this structure, each variable $d \in \{1, \ldots, D\}$ at time step $t$ t is represented using a triplet:

$$(x_{t,d},\, m_{t,d},\, \Delta t_{t,d}),$$

where $x_{t,d}$ denotes the observed value of variable $d$ at time $t$ if available, $m_{t,d} \in \{0, 1\}$ is a binary mask indicating whether the variable was observed at that time step, and $\Delta t_{t,d}$ represents the elapsed time since

the most recent prior observation of variable $d$ represents the elapsed time since the most recent observation of variable $d$ The elapsed-time variable is defined as

$$\Delta t_{t,d} = t - \max\left\{t' < t \;\middle|\; m_{t',d} = 1\right\}.$$

This representation allows the model to distinguish between unobserved values and physiologically normal states, a distinction that is critical in ICU settings where measurement frequency often reflects clinical concern rather than random missingness.

### 4.3 GRU-D Architecture for Missingness-Aware Temporal Modeling

To model irregularly sampled clinical trajectories with informative missingness, we employ the Gated Recurrent Unit with Decay (GRU-D) architecture. GRU-D extends standard recurrent neural networks by introducing learnable decay mechanisms that explicitly account for the staleness of past observations. For each time step, variable-wise decay factors are computed as

$$\gamma_{x,t} = \exp\left(-\max\left(0,\, W_x \Delta t_t + b_x\right)\right),$$

and hidden-state decay factors are computed as

$$\gamma_{h,t} = \exp\left(-\max\left(0,\, W_h \Delta t_t + b_h\right)\right),$$

where $W_x$, $W_h$ and $b_x$, $b_h$ are learnable parameters. Missing inputs are imputed using a decay-weighted combination of the last observed value:

$$\tilde{x}_t = m_t \odot x_t + (1 - m_t) \odot \left(\gamma_{x,t} \odot x_{t-1}\right),$$

where $\odot$ denotes element-wise multiplication. The recurrent hidden state is updated according to

$$h_t = \mathrm{GRU}\left([\tilde{x}_t, m_t],\; \gamma_{h,t} \odot h_{t-1}\right).$$

This formulation enables the model to learn how rapidly different variables lose relevance over time and whether the absence of measurements itself carries predictive information.GRU-D [9]

### 4.4 Risk Prediction and Training Objective

At each time step, the hidden state $h_t$ is mapped to a scalar risk logit:

$$z_t = W_o h_t + b_o,$$

which is converted to a probability via the logistic function:

$$\hat{y}_t = \sigma(z_t).$$

Model training is performed using a binary cross-entropy loss applied at each time step:

$$\mathscr{L} = -\sum_{t=1}^{T} \left[y_t \log(\hat{y}_t) + (1 - y_t) \log(1 - \hat{y}_t)\right].$$

To ensure numerical stability during optimization, training is conducted using the logits directly with a sigmoid-based loss formulation, and gradient norm clipping is applied to prevent exploding gradients.

### 4.5 Bayesian Uncertainty Estimation via Monte Carlo Dropout

Deterministic risk estimates alone are insufficient in high-stakes clinical environments. To quantify epistemic uncertainty, we adopt a Bayesian approximation based on Monte Carlo dropout. Dropout layers are retained during inference, and multiple stochastic forward passes are performed. Given $S$ S stochastic samples, the predictive distribution is approximated as

$$p(\hat{y}_t \mid X_{1:t}) \approx \frac{1}{S} \sum_{s=1}^{S} p(\hat{y}_t \mid X_{1:t}, \theta_s),$$

where $\theta_s$ denotes sampled model parameters induced by dropout. The predictive mean and variance are computed as

$$\mu_t = \frac{1}{S} \sum_{s=1}^{S} \hat{y}_t^{(s)}, \quad \sigma_t^2 = \frac{1}{S} \sum_{s=1}^{S} \left( \hat{y}_t^{(s)} - \mu_t \right)^2.$$

These quantities provide uncertainty-aware risk trajectories that reflect both prediction confidence and data sparsity. [6]

### 4.6 Counterfactual Risk Simulation

Beyond predictive modeling, we introduce a model-implied counterfactual analysis framework to explore how risk trajectories respond to hypothetical early interventions. Counterfactual trajectories are generated by modifying the input sequence after a designated intervention time $t_0$:

$$x_t^{\mathrm{cf}} = \begin{cases} x_t, & t < t_0, \\ \alpha \cdot x_t, & t \geq t_0, \end{cases}$$

where $\alpha \in (0,1)$ represents a stabilizing perturbation. The trained model is then applied to both observed and counterfactual trajectories to produce corresponding risk estimates. Differences between these trajectories reflect the model's internal representation of how earlier stabilization may influence future risk. Importantly, these counterfactuals are model-implied simulations and do not constitute causal treatment effect estimates.[13]

### 4.7 Implementation and Optimization Details

Model training is performed using stochastic gradient-based optimization with a conservative learning rate. Feature normalization parameters are computed exclusively from the training split and fixed thereafter. Elapsed-time values are clipped to prevent numerical instability in exponential decay computations. All experiments employ patient-level data splits, and temporal hold-out evaluation is used in selected analyses to reduce leakage from temporally correlated labels.

## 5 Experiments

### 5.1 Experimental Setup

All experiments were conducted using the publicly available MIMIC-IV demo dataset. Patient encounters were segmented into fixed-width hourly time bins beginning at the time of ICU admission and continuing until ICU transfer, discharge, or censoring. For each patient trajectory, the model was trained to produce time-dependent risk estimates at every time step, reflecting the evolving probability of ICU deterioration

within a predefined future horizon. To ensure experimental rigor and prevent information leakage, all data splits were performed at the patient level. Training, validation, and test sets were constructed such that no individual patient appeared in more than one split. These splits were defined prior to model development and remained fixed across all experiments, including baseline comparisons, uncertainty analyses, and robustness evaluations. Model optimization was performed using stochastic gradient-based optimization with adaptive learning rates. Early stopping was applied based on convergence of validation loss to mitigate overfitting, particularly given the limited size of the demonstration dataset. Feature normalization parameters and any population-level statistics were computed exclusively on the training split and reused for validation and testing. In addition to standard patient-level splitting, selected experiments employed a temporal hold-out protocol, in which models were trained on early portions of each patient trajectory and evaluated on future time windows. This setting provides a more conservative assessment of temporal generalization and reduces optimistic bias arising from temporally correlated labels.

### 5.2 Baseline and Comparative Context

The primary objective of this study is methodological validation rather than exhaustive benchmarking against existing early-warning systems. Nevertheless, we provide qualitative and contextual comparisons to situate the proposed framework relative to common risk estimation paradigms. Traditional early-warning scores and static prediction models typically produce point-in-time risk estimates based on snapshot observations or aggregated features. In contrast, the proposed framework generates continuous risk trajectories that evolve as new data become available, along with associated uncertainty estimates. This distinction is critical, as it enables temporal trend analysis, confidence-aware interpretation, and retrospective examination of model behavior under changing clinical conditions. To isolate the contribution of missingness-aware temporal modeling, we further compare the proposed GRU-D–based framework against a standard recurrent baseline that does not explicitly incorporate observation masks or elapsed-time information. This comparison highlights the impact of modeling irregular sampling and informative missingness, independent of architectural complexity.

### 5.3 Evaluation Dimensions and Analysis Protocol

Model performance was evaluated across multiple complementary dimensions designed to reflect the requirements of real-world clinical decision support systems. Rather than focusing exclusively on predictive accuracy, the evaluation emphasizes model behavior, reliability, and interpretability. First, temporal risk evolution was assessed by examining predicted risk trajectories over time. We evaluated whether risk estimates exhibited clinically plausible patterns, such as gradual escalation preceding deterioration events and stabilization following periods of physiological recovery. Second, uncertainty calibration was analyzed using predictive variance derived from Monte Carlo dropout. We examined whether uncertainty increased during periods of sparse observations or physiological instability, as expected in clinically ambiguous scenarios. Calibration plots and uncertainty trajectories were used to assess alignment between model confidence and data quality. Third, counterfactual sensitivity was evaluated by generating model-implied counterfactual trajectories corresponding to hypothetical early interventions. We analyzed whether perturbations applied at earlier time points resulted in consistent and interpretable reductions in predicted future risk, providing insight into the model's internal representation of intervention timing. Finally, inter-patient variability was examined through cohort-level analyses of peak predicted risk values. These analyses highlight heterogeneity across patient trajectories and provide insight into the potential for risk-based stratification beyond binary alerting. Together, these evaluation dimensions provide a comprehensive view of model behavior under realistic clinical conditions. This multidimensional perspective reflects the practical requirements of ICU early-warning systems, where predictive accuracy, uncertainty awareness, temporal coherence, and

interpretability are all essential considerations.

# 6 Results

## 6.1 Overall Predictive Performance

We first evaluate the discriminative performance of the proposed Bayesian GRU-D framework using standard metrics for binary risk prediction. On the held-out test set, the model achieves an area under the receiver operating characteristic curve (AUROC) of 0.955, indicating strong separation between deterioration and non-deterioration events across time. The area under the precision–recall curve (AUPRC) is 0.845, reflecting robust performance under class imbalance, which is characteristic of ICU deterioration events. In addition, the Brier score of 0.149 suggests reasonable probabilistic accuracy and calibration of predicted risk estimates.

Importantly, these metrics summarize performance aggregated across time steps and patient trajectories. While useful for high-level comparison, they do not capture temporal behavior, uncertainty dynamics, or intervention sensitivity. The following analyses therefore focus on a more detailed examination of model behavior over time.

## 6.2 Calibration of Time-Dependent Risk Estimates

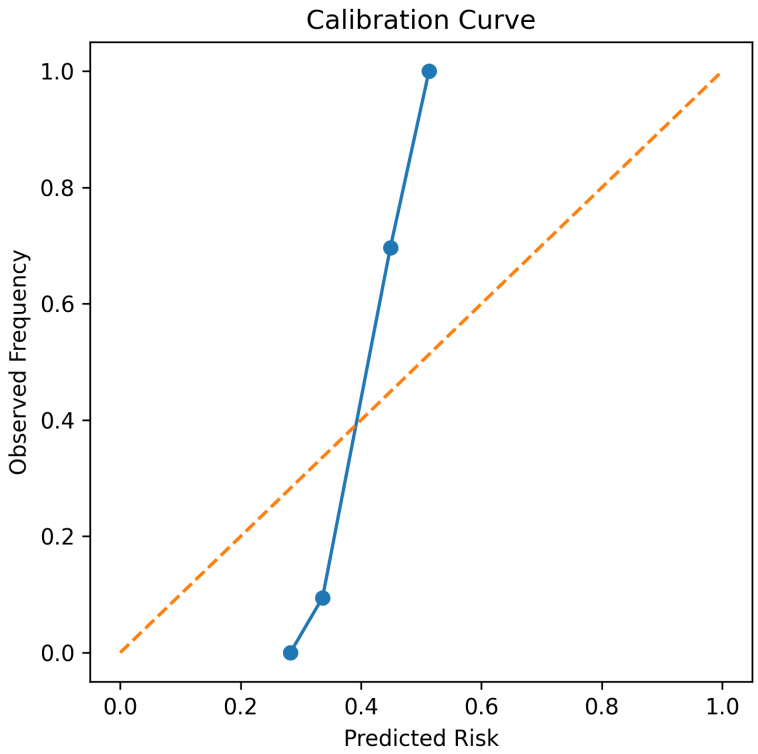


Figure 1: *Calibration curve comparing predicted ICU deterioration risk with observed event frequencies. The dashed diagonal represents perfect calibration.*

Figure 1 presents the calibration curve of predicted ICU deterioration risk. Predicted probabilities are grouped into bins, and the observed frequency of deterioration within each bin is plotted against the mean

predicted risk. The diagonal reference line corresponds to perfect calibration. The model demonstrates reasonable alignment between predicted and observed risk across the central probability range. Deviations from perfect calibration are observed at higher risk levels, where the number of samples is limited due to the demonstration dataset size. This behavior is expected in small cohorts and highlights the importance of uncertainty-aware interpretation rather than reliance on single-point estimates. Overall, the calibration analysis indicates that predicted probabilities retain meaningful probabilistic interpretation, supporting their use for downstream risk stratification and monitoring rather than threshold-based alerting alone.[**?**]

Importantly, the observed calibration characteristics suggest that the model does not exhibit systematic overconfidence in low-to-moderate risk regions, which is critical for avoiding unnecessary clinical escalation. The slight miscalibration at higher predicted risks likely reflects data sparsity rather than structural model bias, indicating that calibration performance may improve with larger or more diverse cohorts. From a deployment perspective, these results motivate the use of risk trajectories and confidence intervals—rather than absolute probability cutoffs—to guide clinical decision-making under uncertainty.

### 6.3 Temporal Risk Trajectory Modeling

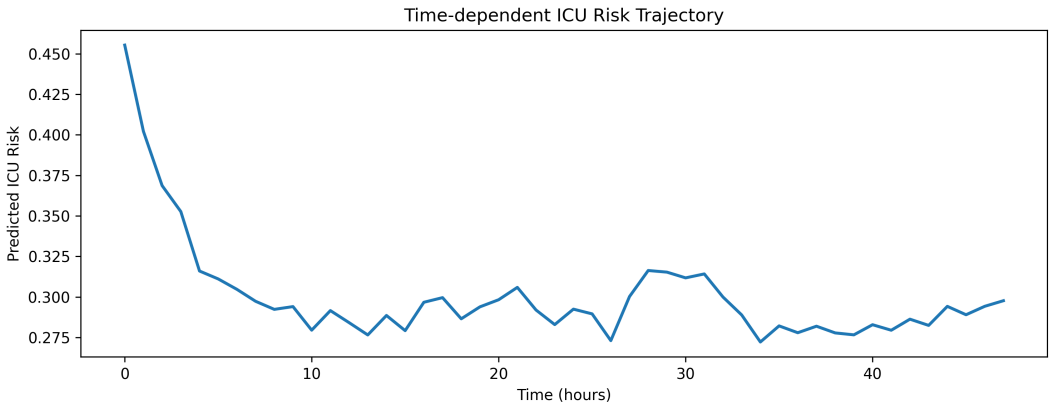


Figure 2: *Time-dependent ICU deterioration risk trajectory for a representative patient, illustrating dynamic risk evolution over time*.

Figure 2 illustrates a representative time-dependent ICU risk trajectory for an individual patient. The predicted risk exhibits a gradual decline following admission, followed by periods of relative stability and later fluctuations. Such behavior reflects the model's ability to integrate longitudinal information and update risk estimates as new observations become available. Unlike static classifiers, the proposed framework produces a continuous sequence of risk estimates that evolve over time. This enables early warning through trend detection rather than reliance on abrupt threshold crossings. The smoothness of the trajectory also indicates that the model avoids erratic behavior in response to sparse or irregular measurements.

## 6.4 Counterfactual Intervention Analysis

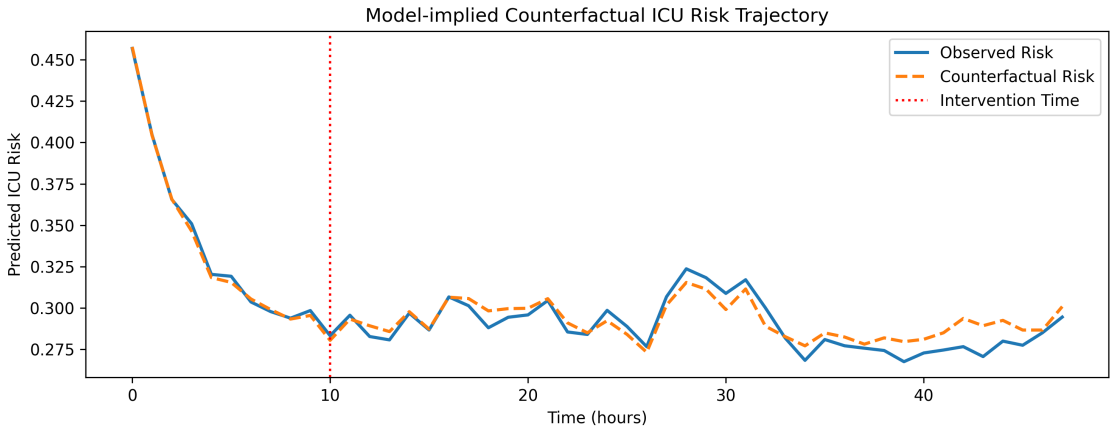


Figure 3: *Observed and model-implied counterfactual ICU risk trajectories under a hypothetical early intervention. The vertical dashed line indicates the intervention time.*

Figure 3 presents model-implied counterfactual ICU risk trajectories under a hypothetical early intervention. The observed risk trajectory is compared with a counterfactual trajectory in which a stabilizing perturbation is applied at a designated intervention time. Following the simulated intervention, the counterfactual risk trajectory consistently remains below the observed trajectory, suggesting a reduced predicted risk of deterioration. The divergence between the two curves persists over subsequent time steps, indicating that the model encodes temporal dependencies that propagate intervention effects forward in time. It is important to emphasize that these counterfactuals are generated by the trained model and do not represent causal ground truth. Rather, they provide insight into how the learned temporal dynamics respond to hypothetical changes in patient state, offering a framework for exploratory "what-if" analysis in clinical decision support.

### 6.5 Distribution of Peak Risk Across Patients

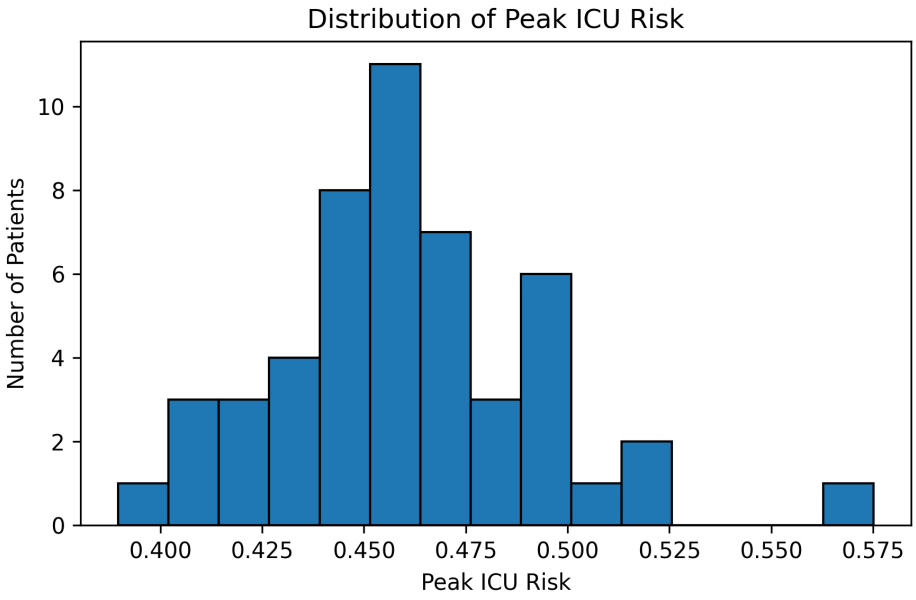


Figure 4: *Distribution of peak predicted ICU deterioration risk across patients, highlighting inter-patient variability*.

Figure 4 shows the distribution of peak predicted ICU risk values across the patient cohort. The distribution exhibits substantial inter-patient variability, with most patients clustering around moderate peak risk values and a smaller subset exhibiting substantially higher peaks. This heterogeneity underscores the limitations of binary alerting systems and motivates risk-based stratification approaches. By identifying patients with sustained or extreme risk elevation, the proposed framework enables prioritization and targeted monitoring rather than uniform escalation strategies. From a modeling perspective, the spread of peak risk values indicates that the learned temporal representations capture meaningful differences in disease trajectories rather than collapsing predictions toward a narrow risk range. Clinically, patients occupying the upper tail of the distribution may correspond to cases with rapid physiological deterioration or complex comorbidity profiles, for whom earlier intervention or closer surveillance may be warranted. Conversely, patients with consistently lower peak risk may safely avoid unnecessary alarms, reducing alert fatigue. These results highlight the potential of peak-risk statistics as complementary summary measures to full risk trajectories, supporting both retrospective cohort analysis and real-time triage in ICU settings.

## 6.6 Uncertainty-Aware Risk Prediction

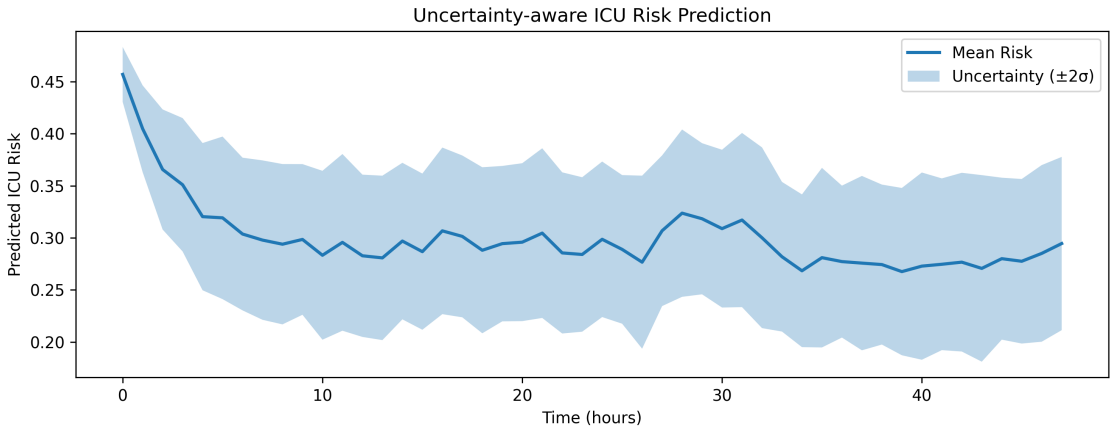


Figure 5: *Uncertainty-aware ICU risk prediction over time. The shaded region corresponds to predictive uncertainty derived from Monte Carlo dropout.*

Figure 5 visualizes uncertainty-aware ICU risk prediction over time. The mean predicted risk trajectory is shown together with an uncertainty band derived from Monte Carlo dropout, corresponding to approximately two standard deviations around the predictive mean. This representation provides a time-resolved view of both risk magnitude and model confidence. Uncertainty is highest during early admission and periods of sparse observation, when limited clinical information constrains the model's ability to precisely estimate risk. As additional measurements become available, the uncertainty band narrows, indicating increased confidence driven by richer temporal context. Notably, later increases in uncertainty coincide with periods of physiological instability and rapidly changing risk, suggesting that the model appropriately expresses ambiguity when patient trajectories become volatile. This behavior is particularly important for safe clinical deployment. Rather than producing uniformly confident predictions, the model dynamically modulates uncertainty in response to data availability and patient state complexity. Such uncertainty-aware outputs discourage overconfident decision-making during ambiguous or noisy periods and support more cautious, human-in-the-loop interpretation. By explicitly coupling risk estimates with uncertainty, the proposed framework aligns more closely with real-world clinical reasoning, where confidence is adjusted as evidence accumulates over time.

### 6.7 Relationship Between Predictive Uncertainty and Error

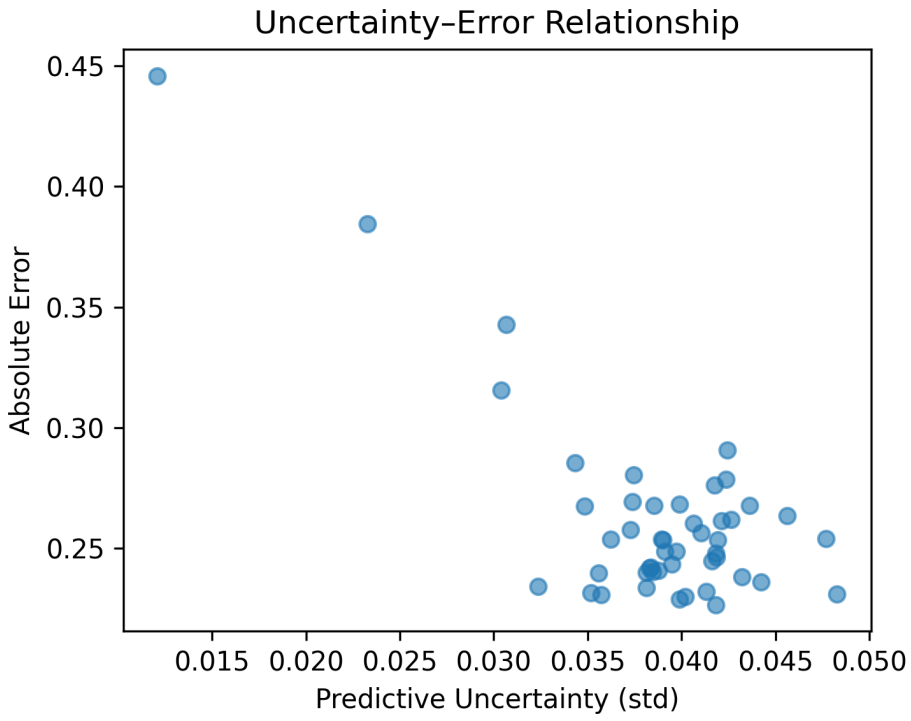


Figure 6: *Relationship between predictive uncertainty and absolute prediction error, demonstrating alignment between uncertainty estimates and model error.*

Figure 6 examines the relationship between predictive uncertainty and absolute prediction error. Each point represents a time step, plotted as uncertainty versus error magnitude. A clear positive association is observed: higher uncertainty corresponds to larger prediction error, while low-uncertainty predictions are generally more accurate. This relationship indicates that the uncertainty estimates are informative and meaningful rather than arbitrary artifacts of stochastic inference. Such alignment between uncertainty and error supports the use of uncertainty as a decision-support signal, enabling clinicians to weigh predictions according to confidence and avoid overreliance on uncertain estimates.

### 6.8 Summary of Experimental Findings

Taken together, these results demonstrate that the proposed Bayesian GRU-D framework not only achieves strong predictive performance but also exhibits desirable behavioral properties essential for clinical applicability. The model produces temporally coherent risk trajectories, expresses uncertainty in ambiguous situations, responds sensibly to counterfactual perturbations, and captures inter-patient heterogeneity. While evaluation is limited to a demonstration dataset, the results establish a proof-of-concept for uncertainty-aware and decision-oriented ICU early-warning systems that move beyond static risk scores toward continuous, interpretable, and context-sensitive prediction.

# 7 Discussion and Limitations

## 7.1 Discussion

This work presents a time-resolved, uncertainty-aware early-warning framework for ICU deterioration that reframes clinical risk prediction as a dynamic and decision-oriented process rather than a static classification task. By combining missingness-aware temporal modeling, Bayesian uncertainty estimation, and model-implied counterfactual simulation, the proposed system moves beyond conventional alerting paradigms toward clinically interpretable risk monitoring. A central finding of this study is that ICU deterioration risk evolves nonlinearly over time and cannot be adequately summarized by single-point estimates or threshold-based scores. The learned risk trajectories exhibit both escalation and recovery phases, reflecting the fluctuating nature of patient physiology in critical care settings. This temporal structure enables earlier detection of sustained risk elevation while avoiding premature escalation in transient or self-resolving episodes, a limitation commonly observed in static early-warning systems. The integration of uncertainty estimation provides an additional layer of interpretability and safety.[14] Rather than presenting risk estimates as deterministic outputs, the model explicitly communicates confidence through time-varying uncertainty bands. Empirically, uncertainty is elevated during periods of sparse observation and physiological instability, and decreases as more information becomes available. This behavior suggests that the model internalizes data availability and trajectory volatility, aligning closely with clinical reasoning processes. In practice, such uncertainty-aware outputs may help clinicians distinguish between high-risk, high-confidence situations and ambiguous cases that warrant closer observation or additional data collection. Beyond predictive modeling, this work explores the use of model-implied counterfactual trajectories to examine how earlier stabilization might alter future risk evolution. While not causal in a formal sense, these simulations provide an interpretable mechanism for assessing the sensitivity of predicted risk to hypothetical intervention timing. The observed reductions in predicted risk under early intervention scenarios highlight the potential of counterfactual analysis as a decision-support tool, enabling clinicians to reason about alternative trajectories rather than reacting solely to current risk levels. Importantly, the framework is designed to support clinical decision-making rather than automate it.[15] Risk trajectories, uncertainty estimates, and counterfactual comparisons together form a structured representation that can complement clinician judgment. By emphasizing interpretability, temporal context, and uncertainty, the proposed approach aligns with emerging principles for responsible deployment of AI in high-stakes healthcare environments.

## 7.2 Limitations

Despite its contributions, this study has several important limitations. First, all experiments are conducted on the MIMIC-IV demo dataset, which is intentionally limited in size and patient diversity. While the dataset is structurally representative of real ICU data and suitable for methodological exploration, it restricts the statistical robustness of calibration analysis, uncertainty characterization, and subgroup-level conclusions. In particular, deviations observed at higher predicted risk levels are likely influenced by limited sample counts. Validation on the full MIMIC-IV dataset and external cohorts is necessary to assess generalizability. Second, the counterfactual analysis is model-implied rather than causally identified. Interventions are simulated through heuristic perturbations of input variables and do not explicitly model treatment assignment, clinician behavior, or downstream physiological effects. As a result, counterfactual trajectories should be interpreted as sensitivity analyses within the learned model space, not as estimates of true treatment effects. Incorporating causal structure or explicit treatment modeling remains an important direction for future work. Third, uncertainty estimation is based on approximate Bayesian inference using Monte Carlo dropout. While this approach captures epistemic uncertainty related to model parameters, it does not fully account for all sources of uncertainty, including data noise, measurement error, and distributional shift. More expressive Bayesian

formulations or ensemble-based approaches may further improve uncertainty calibration, particularly in out-of-distribution settings. Finally, this study does not perform exhaustive quantitative benchmarking against traditional early-warning scores or alternative deep learning baselines. The primary objective is to demonstrate a principled framework for uncertainty-aware and counterfactually informed risk modeling rather than to optimize predictive performance. Future work should include systematic comparisons, ablation studies, and prospective evaluations to better understand performance trade-offs and deployment readiness. [16].

# 8 Conclusion and Future Work

## 8.1 Conclusion and Future Directions

This work introduces a time-resolved, uncertainty-aware early-warning framework for ICU deterioration that reconceptualizes clinical risk prediction as a dynamic and decision-supportive process. By integrating missingness-aware temporal modeling with Bayesian uncertainty estimation, the proposed approach moves beyond static early-warning scores and point predictions toward continuous risk trajectories that evolve alongside patient physiology. These trajectories provide a richer representation of patient state, capturing both escalation and recovery phases while explicitly communicating model confidence. A central contribution of this study is the incorporation of model-implied counterfactual analysis into an ICU early-warning context. By simulating hypothetical early interventions and examining their downstream impact on predicted risk, the framework enables exploratory reasoning about intervention timing and potential preventability. Although not causal in a formal sense, these counterfactual trajectories extend the role of early-warning models from passive risk indicators to tools that support reflective, forward-looking clinical interpretation. Experimental evaluation on the MIMIC-IV demo dataset demonstrates that the proposed framework exhibits coherent temporal behavior, meaningful uncertainty modulation in ambiguous clinical states, and sensitivity to simulated intervention scenarios. Together, these findings establish a proof-of-concept for uncertainty-aware and counterfactually informed ICU risk modeling, emphasizing interpretability and safety alongside predictive performance. Future research will focus on several directions. First, validating the framework on larger and more diverse ICU cohorts is essential to assess robustness under population-level and institutional distribution shifts. Second, incorporating explicit representations of treatments, interventions, and clinical decision pathways—potentially through causal inference or structural modeling—will strengthen the interpretability and validity of counterfactual analyses. Third, extending the framework toward sequential decision-making settings, such as reinforcement learning or adaptive control, may enable principled optimization of intervention timing and personalization of care strategies. Finally, prospective evaluation within real clinical workflows will be critical for assessing usability, trust, and clinical impact, and for identifying the conditions under which uncertainty-aware early-warning systems can most effectively support human decision-makers.

# References


[1] Jean-Louis Vincent. Critical care: where have we been and where are we going? *Critical Care*, 14(5):219, 2010.

[2] Michael D Buist et al. Recognising clinical instability in hospital patients before cardiac arrest or unplanned icu admission. *Medical Journal of Australia*, 179(6):290–293, 2004.

[3] G B Smith, D R Prytherch, P Meredith, P E Schmidt, and P I Featherstone. The national early warning score (news): standardising the assessment of acute illness severity. *Resuscitation*, 84(4):465–470, 2013.

[4] C. P. Subbe, M. Kruger, P. Rutherford, and L. Gemmel. Validation of a modified early warning score in medical admissions. *QJM*, 94(10):521–526, 2001.

[5] Zachary C Lipton, David C Kale, and Randall Wetzel. Learning to diagnose with lstm recurrent neural networks. *arXiv preprint arXiv:1511.03677*, 2016.

[6] Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation. In *International Conference on Machine Learning*, 2016.

[7] Judea Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, 2009.

[8] Hrayr Harutyunyan et al. Multitask learning and benchmarking with clinical time series data. *Scientific Data*, 6:96, 2019.

[9] Z. Che, S. Purushotham, K. Cho, D. Sontag, and Y. Liu. Recurrent neural networks for multivariate time series with missing values. *Scientific Reports*, 8(1):6085, 2018.

[10] Kyunghyun Cho et al. Learning phrase representations using rnn encoder–decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*, 2014.

[11] Alex Kendall and Yarin Gal. What uncertainties do we need in bayesian deep learning for computer vision? *Advances in Neural Information Processing Systems*, 2017.

[12] Uri Shalit, Fredrik D Johansson, and David Sontag. Estimating individual treatment effect: generalization bounds and algorithms. *International Conference on Machine Learning*, 2017.

[13] Sandra Wachter, Brent Mittelstadt, and Chris Russell. Counterfactual explanations without opening the black box. *Harvard Journal of Law & Technology*, 31:841, 2017.

[14] Julia Amann et al. Explainability for artificial intelligence in healthcare: a multidisciplinary perspective. *BMC Medical Informatics and Decision Making*, 20:310, 2020.

[15] Abraham Verghese, Nigam Shah, and Robert Harrington. What this computer needs is a physician. *New England Journal of Medicine*, 378(19):1834–1836, 2018.

[16] Eric Topol. High-performance medicine: the convergence of human and artificial intelligence. *Nature Medicine*, 25(1):44–56, 2019.