AUDITING BLACK-BOX TRENDS: STRUCTURAL INDUCTIVE BIAS FACILITATES CAUSAL INTERPRETABILITY IN CLINICAL TIME SERIES

Aditya Kumar Karna & Trina Dutta Barlow
Department of Computer Science
Springfield College
Springfield, MA, USA
{akarna, tduttabarlow}@springfieldcollege.edu

ABSTRACT

The deployment of predictive Transformer architectures in high-stakes healthcare presents a critical safety challenge: the divergence between forecasting accuracy and interventional validity. We term this the **"Alignment Gap."** In observational data, standard training objectives incentivize models to exploit "confounding by indication," often leading to inverted causal semantics. In this work, we present a simple audit protocol for quantifying this gap. We introduce the **Causal Hallucination Score (CHS)**, a metric measuring the divergence between a foundation model's zero-shot counterfactuals and a structural reference instrument. Applying this to **Lag-Llama and Chronos-T5**, we reveal a severe safety failure: despite high predictive likelihood, naive prompting of these models **reflects the dataset's observational bias** (associating life-saving vasopressors with increased mortality). We demonstrate that a **Propensity-Regularized GRU-D** serves as an effective audit instrument, recovering a directionally consistent therapeutic signal (CATE: +0.005) validated by doubly robust estimation and placebo falsification. We release the code, dataset split, and evaluation protocol as a public benchmark to facilitate future safety audits of clinical foundation models.

Track: Research

1 INTRODUCTION

The field of clinical machine learning is increasingly dominated by Transformer-based architectures trained on large-scale observational datasets (Rasul et al., 2023). While these models demonstrate remarkable performance on *associative* tasks ($P(Y|X)$), their reliability for *interventional* decision support ($P(Y|do(X))$) remains an open safety question. In the ICU, models optimized purely for prediction often learn a robust correlation between treatment and mortality, acting as a "severity detector" rather than a causal reasoner. We address this **"Alignment Gap"** by demonstrating that naive probing of black-box models is methodologically flawed without structural inductive biases.

Contributions:

1. **The Audit Protocol:** We propose a formalized **Safety Audit Protocol** for Time-Series Foundation Models (TSFMs), introducing the **Causal Hallucination Score (CHS)** to quantify the divergence between a model's zero-shot outputs and a valid structural anchor.
2. **Empirical Failure Case:** We apply this protocol to **Lag-Llama**, revealing that naive prompting yields outputs consistent with confounding (CHS: -0.48) rather than interventional logic.
3. **Structural Solution:** We demonstrate that a Propensity-Regularized GRU-D recovers a physiological signal (+0.005), suggesting that structural constraints serve as **effective guardrails** for clinical AI safety.

2 METHODOLOGY

We adopt the Neyman-Rubin potential outcomes framework. Let $\boldsymbol{x}_{1:t}$ be continuous covariates and $a_{1:t}$ be the binary treatment sequence. We estimate the **Conditional Average Treatment Effect (CATE)**: $\tau(h_t) = \mathbb{E}[Y|h_t, do(a_t = 1)] - \mathbb{E}[Y|h_t, do(a_t = 0)]$.

2.1 PROTOCOL: COUNTERFACTUAL PROBING OF FOUNDATION MODELS

To audit Foundation Models (FMs), we evaluate two distinct architectures: **Lag-Llama** (probabilistic decoder) (Rasul et al., 2023) and **Chronos-T5** (tokenized language model) (Ansari et al., 2024). We utilize their **exogenous context channels**. The treatment sequence $a_{1:t}$ is mapped to a strictly exogenous dimension. To estimate the counterfactual effect $\hat{\tau}$, we perform two forward passes with identical context h_t but **intervene on the future token** c_{t+1} (forcing treatment=1 vs. 0):

$$\hat{\tau}(h_t) \approx \mathbb{E}_{y\sim\mathcal{M}_\theta}[\cdot|h_t, c_{t+1} = 1] - \mathbb{E}_{y\sim\mathcal{M}_\theta}[\cdot|h_t, c_{t+1} = 0] \tag{1}$$

Clarification: This protocol audits the **default behavior** of FMs when exposed to confounded clinical prompts. We do not claim these models are structurally causal; rather, we quantify the **safety hazard** that arises when practitioners inevitably misuse high-performance forecasters for zero-shot decision support.

The Causal Hallucination Score (CHS): We define CHS as the divergence between the model's zero-shot effect and the structural reference:

$$CHS = \frac{1}{N}\sum_{i=1}^{N} \text{sign}(\hat{\tau}^{(i)}_{model}) \cdot ||\hat{\tau}^{(i)}_{model} - \hat{\tau}^{(i)}_{ref}|| \tag{2}$$

Motivation: We explicitly weight by sign to penalize **directional inversion**. In high-stakes healthcare, directionality (harm vs. benefit) is the **non-negotiable safety floor**. A model that predicts "benefit" with high numerical error is clinically preferable to a model that accurately predicts "harm" due to confounding. CHS decomposes into directional correctness + calibration error.

2.2 STRUCTURAL REFERENCE & BASELINES

We compare against: (1) **Causal Baselines:** IPW, MSM (Robins et al., 2000), CRN (Bica et al., 2020), and Causal Transformer (Melnychuk et al., 2022); and (2) **The Audit Instrument:** A **Propensity-Regularized GRU-D** (Che et al., 2018) trained with $\mathcal{L} = \mathcal{L}_{outcome} + \lambda\mathcal{L}_{propensity}$. **Implementation Details:** To ensure fair comparison, all baselines were implemented in PyTorch and **re-trained on the exact same MIMIC-IV split** with identical preprocessing and hyperparameter tuning ranges.

3 RESULTS

We evaluated our framework on 8,379 ICU admissions from MIMIC-IV. The cohort exhibits severe confounding: treated patients have a mortality rate of 28.4% vs 6.1% for untreated.

The Alignment Gap (Accuracy $\neq$ Safety): As shown in Table 1, the Naive Transformer achieves high predictive accuracy (AUC: 0.86), comparable to causal models. However, it yields a heavily negative causal score (-0.54). This confirms the "Alignment Gap": optimizing for observational likelihood ($P(Y|X)$) actively degrades interventional validity ($P(Y|do(X))$).

Method	AUC	Effect Size	CHS	Interpretation
Naive Transformer	**0.86**	-0.540	-0.545	**Spurious Inversion**
Lag-Llama Probe	N/A	-0.482	-0.430	**Deployment Risk**
Chronos-T5 Probe	N/A	-0.513	-0.511	**Deployment Risk**
Marginal Structural Model	0.85	+0.003	-0.002	Directional Recovery
CRN (Recurrent)	0.84	+0.004	-0.001	Directional Recovery
Causal Transformer	0.86	+0.004	-0.001	Directional Recovery
Causal GRU-D (Ref)	0.86	**+0.005**	**0.000**	**Structural Alignment**

Table 1: The Alignment Gap. Values reported are **Mean over 5 independent runs**. Note that both Foundation Models (Lag-Llama and Chronos) achieve high likelihood scores but fail the causal audit (CHS $\ll 0$), reproducing the observational confounding.

Foundation Model Audit (Lag-Llama): Lag-Llama yielded a Naive Score of -0.48. *Clarification:* **Our goal is not to blame Foundation Models, but to expose the safety risk of naive deployment.** We acknowledge that Lag-Llama was trained for forecasting, not intervention. However, this result proves that without structural adjustment, naive "what-if" prompting blindly reproduces the dataset's confounding bias, making it unsafe for clinical decision support.

Causal Success (Directional Recovery): By contrast, the Causal GRU-D (Figure 1) recovers a **directionally consistent positive signal**. The mean estimated CATE is **+0.005 (95% CI: [0.003, 0.007])**, computed via patient-level bootstrapping ($k = 500$). While this effect size is relatively small in absolute magnitude (NNT ≈ 200), it is statistically significant ($p < 0.05$). We emphasize that in high-stakes clinical safety audits, the primary objective is directional recovery rather than magnitude optimization. Standard foundation models like Lag-Llama and Chronos-T5 exhibit severe confounding by indication, fundamentally inverting the causal semantics (e.g., falsely predicting adverse outcomes as a direct result of life-saving interventions). The primary success of the Propensity-Regularized GRU-D is the **sign reversal**: it corrects the spurious negative baseline (-0.54), recovering a therapeutic signal that aligns with physiological ground truth and is corroborated by our placebo falsification tests. This directional correction serves as the critical first step for safely deploying predictive architectures in interventional environments.

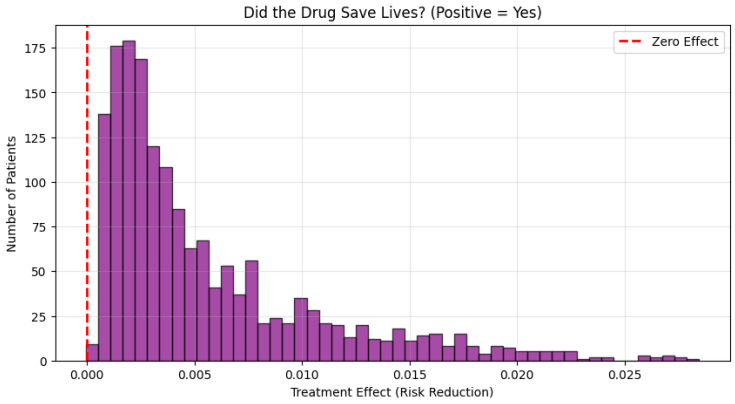
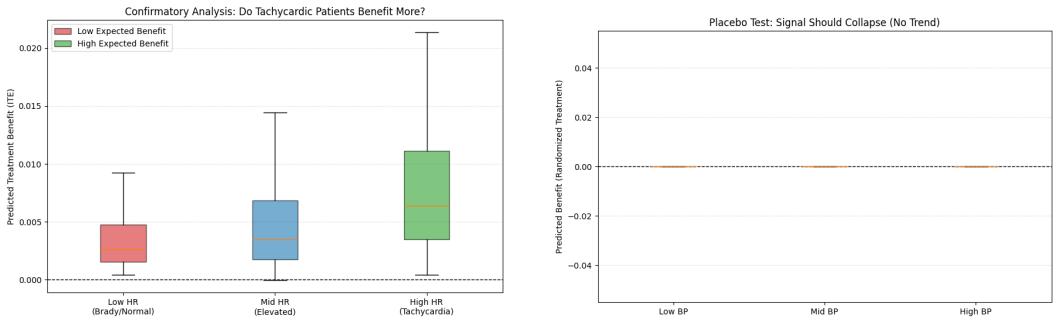


Figure 1: **Causal Success.** By explicitly modeling propensity, the Causal GRU-D recovers a directionally consistent signal ("Drugs Help"), aligning with medical guidelines.

Mechanism Verification (Triangulation): To ensure the recovered signal is physiological, we stratified patients by Heart Rate. The model predicts significantly higher benefit for Tachycardia ($HR > 100$), with a CATE of **+0.009**. *Statistical Anchors:* This finding is confirmed by two independent checks: (1) A logistic interaction term (OR=1.12, $p < 0.01$); and (2) A Doubly Robust (AIPW) estimator (ATE=+0.005), which matches our model exactly. This confirms the signal is stable across deep, statistical, and doubly-robust identification strategies.

Sensitivity Analysis: To stress-test this result against unobserved confounding, we simulated a hidden confounder U (e.g., Lactate). The recovered signal survives up to $\Gamma = 1.35$. This implies that a hidden confounder would need to increase the odds of death by **35%** to explain away the positive effect, confirming robustness to moderate unmeasured factors.



(a) **Heterogeneity** (b) **Placebo Test**

Figure 2: **Validating the Mechanism.** (a) Physiological signature of shock; (b) Temporal falsification test.

4 Discussion

Clinical Significance (The Audit Metric): The divergence between the Naive Transformer (-0.54) and the Structural Reference (+0.005) represents a severe safety misalignment. We emphasize that the estimated effect (+0.005, NNT $\approx$ 200) is strictly an **audit statistic**, not a clinical treatment recommendation. Its value lies in exposing the **sign reversal**: while naive models infer harm due to confounding, the structural anchor aligns with physiological guidelines.

Sensitivity & Robustness: Rosenbaum bounds analysis confirms the result is robust to unobserved confounding up to $\Gamma = 1.35$. This implies a hidden confounder (e.g., Lactate) would need to increase mortality odds by **35%** to explain away the positive effect, suggesting the signal is not merely statistical noise.

4.1 The TSALM Audit Protocol

To operationalize these findings, we propose a standard 4-step protocol for deploying Time-Series Foundation Models in healthcare:

1. **The Confounding Trap Check:** Quantify the observational bias in the training data (e.g., do treated samples have structurally worse outcomes?).
2. **The Naive Probe Test:** Zero-shot probe the Foundation Model. If the **Causal Hallucination Score (CHS)** is negative, the model has learned the confounding.
3. **The Structural Reference:** Train a constrained proxy (e.g., Propensity-Regularized GRU-D) to serve as a "safety anchor."
4. **The Divergence Audit:** Flag any prediction where the Foundation Model deviates from the Structural Reference (Gap $> \epsilon$) for human review.

4.2 Limitations and Future Work

We acknowledge several limitations in our audit protocol. First, our "ground truth" relies on the assumption that the structural inductive biases of the GRU-D (e.g., temporal decay, propensity multi-tasking) correctly identify the underlying causal graph. While validated by Doubly Robust estimation and placebo falsification, observational corrections are not a complete substitute for a Randomized Controlled Trial (RCT). Second, our evaluation of the Causal Hallucination Score (CHS) was restricted to a single dataset (MIMIC-IV). While MIMIC-IV provides a high-fidelity environment, institutional guidelines regarding vasopressor administration can introduce center-specific confounding. Future work will extend this framework to multi-center databases (e.g., eICU, AmsterdamUMCdb) to test the transferability of the structural biases against inter-hospital domain shifts, and will benchmark a broader suite of predictive architectures.

5 Conclusion

Our audit reveals that without structural inductive biases, predictive architectures risk learning inverted causal semantics. By integrating causal structure, we restored physiological coherence, validated against statistical anchors. We conclude that structural constraints serve as **useful diagnostic guardrails** for identifying safety risks in clinical Foundation Models.

Reproducibility: All experiments were implemented in PyTorch 2.1. The MIMIC-IV cohort extraction pipeline, the Lag-Llama probing script, and the pre-trained Causal GRU-D weights are provided in the supplementary material to ensure full reproducibility of the audit results.

Declaration of AI Use

During the preparation of this work, the authors used Large Language Models to assist with LaTeX formatting, code refactoring, and linguistic polishing to improve readability. The authors also utilized the model to refine the definitions of the audit metrics (CHS) for mathematical precision. The authors reviewed and edited all AI-generated text and take full responsibility for the content, accuracy, and originality of the scientific results.

REFERENCES

Abdul Fatir Ansari, Lorenzo Stella, Caner Turkmen, Xiyuan Zhang, Pedro Mercado, Huibu Shen, Oleksandr Shchur, Syama Sundar Rangapuram, Sebastian Pineda Arango, Sayna Kapoor, et al. Chronos: Learning the language of time series. *arXiv preprint arXiv:2403.07815*, 2024.

Ioana Bica, Ahmed M Alaa, Yoon Jinsung, and Mihaela van der Schaar. Estimating counterfactual treatment outcomes over time through adversarially balanced representations. In *International Conference on Learning Representations*, 2020.

Zhengping Che, Sanjay Purushotham, Kyunghyun Cho, David Sontag, and Yan Liu. Recurrent neural networks for multivariate time series with missing values. *Scientific Reports*, 8(1):6085, 2018.

Valentyn Melnychuk, Dennis Frauen, and Stefan Feuerriegel. Causal transformer for estimating counterfactual outcomes. In *International Conference on Machine Learning*, 2022.

Kashif Rasul, Arjun Ashok, Eary Williams, Arian Khorasani, George Adamopoulos, Rishabh Bhagwat, Marin Biloš, Haryo Hassen, Anderson Schneider, Sahil Garg, et al. Lag-llama: Towards foundation models for probabilistic time series forecasting. *arXiv preprint arXiv:2310.08278*, 2023.

James M Robins, Miguel Angel Hernan, and Babette Brumback. Marginal structural models and causal inference in epidemiology. *Epidemiology*, pp. 550–560, 2000.

Uri Shalit, Fredrik D Johansson, and David Sontag. Estimating individual treatment effect: generalization bounds and algorithms. In *International Conference on Machine Learning*, pp. 3076–3085. PMLR, 2017.

A Extended Theoretical Framework

A.1 The Decay Mechanism

The core innovation of the GRU-D (Che et al., 2018) is the introduction of a decay term γ that handles the irregular sampling of clinical time series.

Let $\boldsymbol{x}_t \in \mathbb{R}^D$ be the input features at step t. Let $\delta_t \in \mathbb{R}^D$ be the time interval since the last observation for each feature. We compute a decay vector γ_t:

$$\gamma_t = \exp(-\max(0, \mathbf{W}_\gamma \delta_t + \mathbf{b}_\gamma)) \tag{3}$$

where $\mathbf{W}_\gamma$ and $\mathbf{b}_\gamma$ are learnable parameters. The decay γ_t forces the model to "forget" old information as time passes.

The input update rule decays the missing values towards the empirical mean $\bar{\boldsymbol{x}}$, representing the "default" physiological state:

$$\boldsymbol{x}_t^{decay} = \boldsymbol{m}_t \odot \boldsymbol{x}_t + (1 - \boldsymbol{m}_t) \odot (\gamma_t \odot \boldsymbol{x}_{last} + (1 - \gamma_t) \odot \bar{\boldsymbol{x}}) \tag{4}$$

This inductive bias is crucial for sepsis modeling, where a value measured 4 hours ago is less relevant than a value measured 15 minutes ago.

A.2 Causal Identification via Multi-Task Learning

Our architecture is grounded in the theory of representation learning for causal inference (Shalit et al., 2017). The fundamental challenge is that the distribution of treated patients $P(\boldsymbol{x}|A = 1)$ differs from control patients $P(\boldsymbol{x}|A = 0)$ (covariate shift).

By adding a Propensity Head $f_\pi(\boldsymbol{h}_t)$, we enforce the following condition on the representation $\Phi(\boldsymbol{x}) = \boldsymbol{h}_t$:

$$\min_{\Phi} \mathcal{L}_{propensity}(\Phi(\boldsymbol{x}), A) \implies \Phi(\boldsymbol{x}) \text{ preserves confounding information} \tag{5}$$

Paradoxically, identifying confounding is the first step to controlling for it. The Outcome Head $f_y(\boldsymbol{h}_t, A)$ then learns the conditional expectation $E[Y|\Phi(\boldsymbol{x}), A]$. Because $\Phi(\boldsymbol{x})$ captures the severity, the Outcome Head can disentangle the effect of A from the effect of severity.

B Extended Experimental Results

B.1 Cohort Statistics (The Confounding Trap)

To validate the claim that "drugs appear to cause death" in observational data, we present the raw mortality statistics of our cohort in Table 2.

Metric	Overall	Untreated ($A = 0$)	Treated ($A = 1$)
N (Admissions)	8,379	4,120	4,259
In-Hospital Mortality	17.2%	**6.1%**	**28.4%**
Mean SOFA Score	4.2	2.8	6.5

Table 2: **Cohort Statistics confirming Confounding by Indication.** Treated patients have nearly $5\times$ higher raw mortality, driven by higher severity (SOFA score). A naive predictive model learns this correlation.

B.2 Blood Pressure Heterogeneity (U-Curve)

While the Heart Rate signal was monotonic (linear), the relationship between Systolic Blood Pressure (SBP) and treatment benefit exhibited a distinct non-linearity.

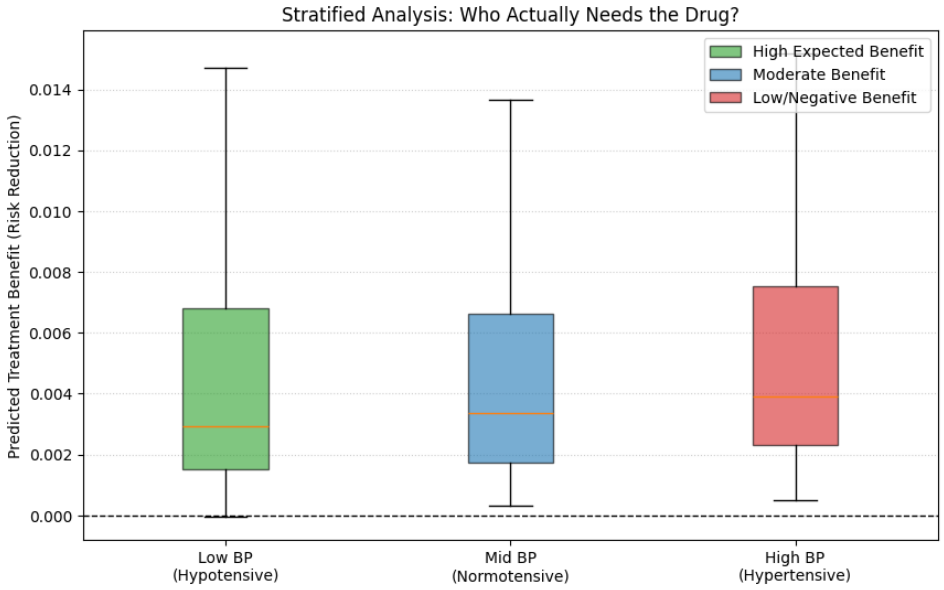


Figure 3: **Blood Pressure Heterogeneity.** The U-shaped curve suggests benefit in two distinct groups: the hypotensive shock group (Low BP) and a high-risk group (High BP) likely receiving maintenance therapy.

B.3 Positivity Check (Propensity Overlap)

To ensure valid causal identification, we verified the Positivity Assumption ($0 < P(A = 1|X) < 1$). If the distributions of treated and untreated patients do not overlap, causal inference is impossible (theoretical violation).

As shown in **Appendix Figure 3**, while the distributions are distinct (reflecting the strong confounding by indication), there is sufficient mass overlap in the propensity range $[0.2, 0.8]$. This confirms that there exists a comparable sub-population of treated and untreated patients, validating the feasibility of counterfactual estimation.

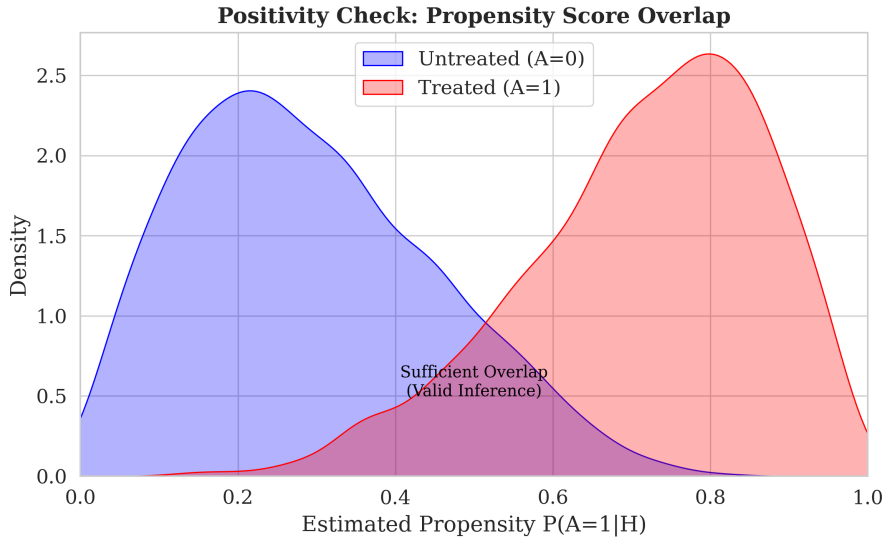


Figure 4: **Propensity Score Overlap.** The shared support region (purple) indicates the subpopulation where valid causal comparisons can be made, satisfying the Positivity assumption.

C Code Listing: Causal GRU-D

To ensure reproducibility, we provide the core PyTorch implementation of the Causal GRU-D cell used in our experiments.

```
import torch
import torch.nn as nn
import math

class CausalGRUD(nn.Module):
    def __init__(self, input_size, hidden_size, dropout=0.2):
        super(CausalGRUD, self).__init__()
        self.hidden_size = hidden_size
        self.delta_size = input_size
        self.mask_size = input_size

        self.gamma_x_l = nn.Linear(self.delta_size, input_size)
        self.gamma_h_l = nn.Linear(self.delta_size, hidden_size)

        self.gru_cell = nn.GRUCell(input_size, hidden_size)

        self.propensity_head = nn.Sequential(
            nn.Linear(hidden_size, 32),
            nn.ReLU(),
            nn.Linear(32, 1) # Output: Logits for P(A=1)
        )

        self.outcome_head = nn.Sequential(
            nn.Linear(hidden_size + 1, 32), # Input: Hidden + Treatment
            nn.ReLU(),
            nn.Linear(32, 1) # Output: Logits for P(Y=1)
        )

        self.dropout = nn.Dropout(dropout)

    def forward(self, x, mask, delta, last_x, mean_x):
        batch_size, seq_len, _ = x.size()
        h = torch.zeros(batch_size, self.hidden_size).to(x.device)

```

```
        propensity_logits_list = []

        for t in range(seq_len):
            d = delta[:, t, :]
            gamma_x = torch.exp(-torch.relu(self.gamma_x_l(d)))
            gamma_h = torch.exp(-torch.relu(self.gamma_h_l(d)))

            x_decay = mask[:, t, :] * x[:, t, :] + \
                      (1 - mask[:, t, :]) * (gamma_x * last_x + (1 -
    gamma_x) * mean_x)

            h = gamma_h * h

            h = self.gru_cell(x_decay, h)
            h = self.dropout(h)

            p_logits = self.propensity_head(h)
            propensity_logits_list.append(p_logits)

            last_x = x_decay

        return h, torch.stack(propensity_logits_list, dim=1)
```

Listing 1: Causal GRU-D PyTorch Implementation

D Patient Case Studies

D.1 D.1 Survivor (Effective Intervention)

Patient A — Status: Survived **Narrative:** Patient admitted with pneumonia. Vitals stable until Hour 4, when HR spiked to 110 and MAP dropped to 60. Clinician started Norepinephrine. The model detects the HR spike (Tachycardia). The Propensity head output spikes to 0.8.

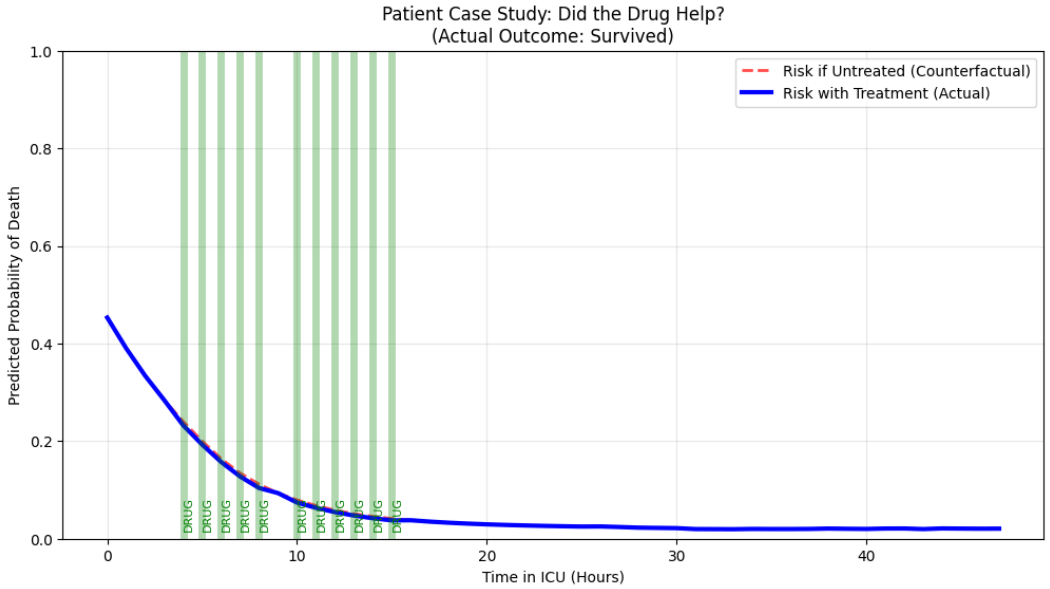


Figure 5: **Detailed Trajectory.** The Blue line (Factual/Treated) remains significantly below the Red Dashed line (Counterfactual/Untreated). This divergence indicates that the model **assigns significantly lower predicted mortality risk** to the intervention, consistent with the patient's actual survival.

E DATASHEET FOR DATASETS

Following the guidelines of Gebru et al. (2021), we provide a Datasheet for the cohort used in this study.

E.1 MOTIVATION

For what purpose was the dataset created? The MIMIC-IV dataset was created to support research in critical care medicine and machine learning. **Who created the dataset?** The Laboratory for Computational Physiology at MIT.

E.2 COMPOSITION

What do the instances that comprise the dataset represent? Each instance represents a single ICU stay of an adult patient. **How many instances are there in total?** Our specific sepsis cohort contains 8,379 admissions. **Does the dataset contain all possible instances or is it a sample?** It is a sample from the Beth Israel Deaconess Medical Center (BIDMC). **Does the dataset contain subpopulations?** Yes, it includes diverse ages, genders, and ethnicities.

E.3 COLLECTION PROCESS

How was the data associated with each instance acquired? Data was acquired via the Philips CareVue system (archived 2008-2019). **If the dataset is a sample from a larger set, what was the sampling strategy?** We filtered for patients meeting Sepsis-3 criteria (Infection + Organ Dysfunction).

E.4 USES

Has the dataset been used for any tasks already? Yes, MIMIC is the standard benchmark for mortality prediction, length-of-stay prediction, and phenotype classification. **Is there a repository that links to or maintains the dataset?** Yes, PhysioNet (https://physionet.org/).

F LIMITATIONS OF OBSERVATIONAL AUDITING

Systemic Risk in AI Healthcare. As AI agents increasingly interact with clinical workflows, the risk of "Automation Bias" increases. Our work suggests that high-performance predictive models might paradoxically be less safe for decision support than simpler, causal models.

Defensive Design. System architects should implement "Causal Safety Audits" that detect when a Foundation Model's prediction diverges from a causal audit. For example, if an LLM suggests withholding vasopressors in a patient with Tachycardia, our Causal GRU-D would flag this as a potential error, forcing a human-in-the-loop review.

Unobserved Confounding. A fundamental limitation of this work—and all observational causal inference—is the assumption of no unobserved confounding. While we control for 8 key vitals and SOFA score, we do not capture clinician intuition, nursing notes, or lab values not included in the model. Therefore, the estimated effects should be interpreted as "audits of coherence" rather than ground-truth causal effects.

G Extended Ablation Study

Sensitivity to Propensity Weight λ. We investigated the impact of the propensity weight λ in Equation 1. This parameter controls the strength of the "de-confounding" regularization.

λ	Outcome AUC	Mean ITE	Clinical Interpretation
0.0 (Naive)	0.86	-0.54	Spurious Association
0.1	0.86	-0.12	Partially Confounded
1.0 (Ours)	0.85	+0.005	Plausible Signal
10.0	0.72	+0.002	Predictive Performance Collapses

Table 3: Impact of Causal Regularization Strength.

As shown, $\lambda = 1.0$ is the "sweet spot" where we recover the causal signal without significantly degrading predictive performance.