Developing Bias Identification and Mitigation Techniques for Clinical Prediction Models

Hannah Liu*
Dept. of Electrical and Electronic Engineering
Imperial College London
hannah.liu25@imperial.ac.uk

Lucie Pasquier*
Dept. of Electrical and Electronic Engineering
Imperial College London
lucie.pasquier25@imperial.ac.uk

Abstract

The effects of missing data have been studied primarily in the context of prediction, where the literature has affirmed that informative missingness can be a strong predictive signal in clinical settings. However, when this signal interacts with deployment scenarios, it leads to consequential outcomes. We present a paper showing that clinical prediction models (CPMs) can learn to use missing data as a predictive signal, thereby creating an undesirable feedback loop. We also introduce two mitigation methods that resolve these issues: an uncertainty-triggered measurement intervention and a causal RLHF pipeline. Our bias identification methods confirmed that models trained on biased measurement data learn to treat a lack of data as a proxy for health, and our bias mitigation methods were effective in reducing bias relative to the baseline model.

1 Introduction

The integration of artificial intelligence (AI) into the medical field raises concerns about how bias can emerge, compound, and cause significant harm in clinical processes [14, 5, 16, 2]. One such concern is missing data, which occurs when clinical data generation is driven by decisions about whom to test and monitor. These decisions are often influenced by both clinical and non-clinical factors [14]. As a result, electronic health record data is inherently selective and often incomplete. However, when certain patients are measured less frequently, missingness itself can become informative. Prior work shows that models may use the presence or absence of measurements as a proxy for risk, effectively learning that unmeasured patients are lower risk [17]. While this can improve retrospective performance, it raises concerns at deployment. If model predictions influence future measurement decisions, the data-generating process becomes coupled with the model. This can induce a feedback loop: less measurement leads to sparser data, lower predicted risk, and consequently even less measurement. Over time, such dynamics may compound disparities across patient groups.

Despite extensive work on handling missing data [4, 6, 18], little attention has been paid to how missingness-based signals behave under deployment. In this work, we study how selective measurement can induce feedback loops in clinical prediction systems and when these dynamics amplify bias. Our contributions include: 1) demonstrating that CPMs can learn to use missing data as a predictive signal, which creates an undesirable feedback loop, and 2) introducing mitigation methods that fix these observed issues.

*Equal contribution.

[1]Code: https://github.com/lucie-pasquier/bias-identification-mitigation-for-cpms

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

2 Related work

2.1 Informative missingness

Informative missingness is a concept that treats missing data in electronic health records (EHRs) as a good predictive signal rather than a nuisance is called [21, 19, 20, 3]. For example, clinicians who selectively order tests for patients create informative patterns that reflect aspects of medical decision-making. Informative missingness is prevalent in clinical predictive settings because data collection is driven by clinical requirements of patients and clinicians rather than by a fixed measurement protocol [19]. However, because of this, some patients may be unmeasured if internal patient or clinician decisions deemed certain diagnostic processes unnecessary, thus creating space for selectively biased data to be produced. One particular example is the selective labels problem, in which outcomes are observed only for patients that a clinician chooses to test. As a result, the model learns $P(Y \mid X, D{=}1)$ rather than $P(Y \mid X)$, making standard evaluation unreliable (Y is the true label, X is the covariates available for prediction, and D is the human decision) [12, 7].

This literature establishes that informative missingness provides a predictive signal that can be beneficial. However, [12, 7] shows that more work is needed to examine what happens when a deployed model's predictions feed back into the measurement process that generates the predictive signal. This is the interaction we seek to investigate.

2.2 Performative prediction

Performative prediction describes settings in which a model's predictions have downstream effects on clinical decisions and care pathways that determine the outcomes the model aims to predict [15]. Formally, we can represent this performative risk minimization problem as $\mathrm{PR}(\theta) \triangleq \mathop{\mathbb{E}}_{Z \sim \mathcal{D}(\theta)} \ell(Z; \theta)$, where the data distribution $\mathcal{D}$ itself depends on model parameters.

Existing literature has shown how individual feedback loops in ML decision-making perpetuate this concept, creating historical bias because decisions that alter current inherent properties can affect future features [13, 1, 8]. However, this prior work has not necessarily examined how this mechanism interacts with informative missingness, in which the measurement pattern is used as a predictive feature. We address this gap by showing that informative missingness and performative prediction interact to create an undesirable feedback loop and by proposing mitigation methods that break this bias cycle.

3 Preliminaries

In forming our bias mitigation techniques, we decide to translate an RLHF pipeline into a causal graph, as shown in Figure A3. However, we notice that the resulting graph creates a cycle between the model, the output, and the reward model. As a result, it breaks the directed acyclic graph (DAG) assumption that underlies standard causal inference:

> Let $\mathcal{G} = (V, E)$ be the causal graph. Standard causal inference methods, including d-separation, require $\mathcal{G}$ to be a directed acyclic graph (DAG), i.e. there exists no directed path $v_1 \to v_2 \to \cdots \to v_k \to v_1$ for any $\{v_1, \ldots, v_k\} \subseteq V$.

Our analysis of performative prediction shows that these feedback loops and cycles can create downstream effects on the outcomes themselves. Therefore, having a bias mitigation technique in this structure would be counterproductive.

We resolve this issue by splitting the original causal graph (see Figure A3) into three piecewise causal chains that run in sequential time steps. This separation can be seen in Figure A4. Our goal was to apply D-separation to each of these causal chains to determine conditional independence, but we encountered an issue: a node observed in one time step must be unobserved in the next. For example, $x \to M \to y$ requires M to be observed for conditional independence to occur, but in $M \to y \to RM$, M must be unobserved. Resolving these observability issues was critical to determining our solution. This split also does not provide an avenue for us to consider potential hidden confounders, one of which could be the influence of missing data via informative missingness.

To resolve these challenges, we assume the following when building our causal RLHF model:

1. **Confounder containment.** All hidden confounders are absorbed into a single latent node h. That is, for any unobserved variable U that is a common cause of two or more nodes in the graph, U is contained in h. Formally, let $V_{\text{obs}} = \{x, M_1, M_2, y_1, y_2, RM_1, RM_2\}$ denote the set of observed nodes. For any pair $V_i, V_j \in V_{\text{obs}}$, there exists no latent common cause outside of h:
$$\nexists\, U \notin V_{\text{obs}} \cup \{h\} \;\text{ such that }\; U \to V_i \;\text{ and }\; U \to V_j. \tag{1}$$
This ensures that conditioning on h is sufficient to block all backdoor paths through unobserved confounders, enabling valid causal identification within each piecewise subgraph.
2. **Temporal copy exclusivity.** Nodes that appear at multiple timesteps are copies of the same underlying entity unrolled over time. Specifically, M_1 and M_2 are instantiations of the same model M at consecutive timesteps, and similarly, y_1 and y_2 for the output and RM_1 and RM_2 for the reward model. At any given timestep, only one copy is observed while the other is unobserved:
$$\begin{aligned} o(M_1) = 1 \implies o(M_2) = 0, \\ o(M_2) = 1 \implies o(M_1) = 0, \end{aligned} \tag{2}$$
where $o(\cdot) \in \{0, 1\}$ denotes the observability indicator. The same exclusivity holds for (y_1, y_2) and (RM_1, RM_2). This temporal unrolling serves two purposes: it breaks the cycle in the original causal graph, and it encodes the fact that the system cannot simultaneously be in two stages of the feedback loop.

4 Methods

4.1 Bias identification

We propose a two-task methodology for identifying observation-driven bias in clinical risk scoring systems, validated in both a controlled synthetic environment and a real-world clinical dataset.

Task 1 asks whether a model trained on informatively missing data encodes an *observability shortcut*: treating unmeasuredness as a signal of lower clinical risk. We inspect learned coefficients directly and develop a counterfactual masking audit that quantifies how much predicted risk drops when a patient is artificially treated as unmeasured.

Task 2 asks whether this shortcut, once encoded, creates a self-sustaining feedback loop under deployment. We simulate a risk-driven monitoring policy and track whether disparities in measurement rate, predicted risk, and gap length emerge and persist despite equal underlying severity.

Together, Task 1 establishes that the bias exists within the model and Task 2 establishes that deployment actively compounds it. Both tasks are evaluated on the synthetic experiment first, then applied to MIMIC-IV to assess whether the same signatures appear in real hospital data.

4.1.1 Task 1: observability audit

The observability audit is a two-stage procedure for detecting and quantifying whether a model has encoded measurement patterns as proxies for clinical risk.

Stage 1: Coefficient inspection. We train a logistic regression model to predict outcome Y from $(\tilde{x}_t, m_t, \delta_t, c_t)$. The observability shortcut is directly detectable in the learned weights: negative coefficients on m_t and δ_t indicate the model predicts lower risk for unmeasured patients, and as gaps accumulate. We assess statistical significance via Wald tests, reporting coefficients, z-scores, p-values, and 95% confidence intervals. The shortcut is confirmed if m_t and δ_t carry significant negative coefficients dominant in magnitude over $\tilde{x}_t$. To verify the shortcut is a structural property of the data rather than an artifact of model choice, we additionally train an MLP (64 units, ReLU) and XGBoost on identical splits, using SHAP-based feature attribution to assess whether measurement pattern features remain dominant across model classes.

Stage 2: counterfactual masking audit. To quantify the practical consequence of the shortcut on individual predictions, for each held-out test record, we compute:
$$\Delta\text{risk} = f(\tilde{x}_t, m_t, \delta_t, c_t) - f(\tilde{x}_t,\ 1,\ \delta_t + 1,\ c_t) \tag{3}$$

where $f(\cdot)$ denotes predicted probability. This measures how much predicted risk drops when we counterfactually set $m_t = 1$ and increment δ_t by one, the two features that would differ if no measurement had occurred. A positive mean Δrisk confirms the shortcut operates on real predictions. We report the mean Δrisk by group to assess effects across logistic regression, MLP, and XGBoost.

4.1.2 Task 2: score-based simulation

The score-based simulation deploys the trained model as a monitoring policy to test whether measurement disparities emerge and persist over time, with no new external bias introduced after training.

Initialization. Each patient is initialized from their actual end-of-training feature state, preserving the accumulated measurement disadvantage that Group 1 patients carry into deployment.

Simulation procedure. We simulate $T = 30$ deployment timesteps. At each timestep t, for every patient i:

1. Predicted risk is computed: $\hat{p}_i^t = f(\text{features}_i^t)$
2. Measurement probability is assigned:
$$\pi_i^t = \begin{cases} p_{\text{high}} = 0.85 & \text{if } \hat{p}_i^t > \tau = 0.5 \\ p_{\text{low}} = 0.15 & \text{otherwise} \end{cases} \tag{4}$$
3. Measurement is drawn: $M_i^t \sim \text{Bernoulli}(\pi_i^t)$
4. Features update: if measured, gap resets, $\tilde{x}_t$ refreshes, count increments; otherwise gap increments, $\tilde{x}_t$ carries forward.
5. True severity transitions via the Markov chain, independently of measurement.

Tracked metrics. We record mean measurement rate, predicted risk, gap length, and true severity prevalence by group at each timestep. True severity serves as a control: any disparity in the remaining metrics is attributable to the model and policy rather than genuine clinical differences. Disparity is reported as Group 0 minus Group 1 averaged across the deployment period. The model is frozen throughout to isolate the feedback loop from the compounding effects of retraining, which we acknowledge as a limitation.

4.2 Bias mitigation

We propose two bias mitigation techniques to reduce bias in our baseline model.

Task 3 proposes an uncertainty-triggered measurement intervention, which assigns the patient a higher measurement probability during the score-based policy simulation phase (Task 2) if two specific thresholds are exceeded.

Task 4 proposes a causal RLHF pipeline that offers an alternative to the typical RLHF processes commonly found in clinical settings. It corrects the biased predictions directly, producing debiased scores that naturally equalize monitoring rates when fed into the same threshold policy.

4.2.1 Task 3: uncertainty-triggered measurement

Method. We augment the baseline policy with two additional triggers. A patient is assigned p_{high} if either their predictive uncertainty exceeds a threshold, $\hat{p}_i^t(1 - \hat{p}_i^t) > u_0$, or their measurement gap exceeds a maximum allowed duration, $\delta_i^t > g_0$. The uncertainty criterion targets patients for whom the model lacks confidence, while the gap criterion directly addresses chronically under-observed patients regardless of predicted risk. Together, they break the core feedback mechanism: a patient can no longer be systematically under-measured simply because the model is confident they are low risk.

Justification of parameters. We set $u_0 = 0.20$, corresponding to 80% of the maximum Bernoulli variance (0.25 at $\hat{p} = 0.5$), and $g_0 = 3$ timesteps, a gap duration that is routine for the under-measured group, but rare for the reference group in our synthetic data.

4.2.2 Task 4: causal RLHF

Causal graph. Figure 1 shows the final construction of our causal RLHF pipeline after implementing the assumptions made in our Preliminaries section. The splitting of nodes observed more than once

across timesteps resolves the violation of the DAG assumption and the observability conflicts. The addition of the latent node h resolves the issue of not accounting for hidden confounders.

Components. M_1 is frozen throughout, analogous to a pre-trained base model in standard RLHF. M_2 maps biased predictions and observable proxies to debiased targets: $y_2 = M_2(y_1,\ \delta_t,\ m_t,\ y_1 \cdot \delta_t,\ y_1 \cdot m_t)$. The interaction terms capture how gap-induced underestimation scales with predicted risk. M_2 learns only the debiasing correction from observable proxies and does not observe group membership. To prevent abrupt parameter jumps, M_2 blends each round's predictions with the previous round's: $y_2^{(\text{blend})} = (1-\eta)\, y_2^{(r-1)} + \eta\, y_2^{(r,\text{new})}$.

RM_2 produces correction targets via *counterfactual gap correction*: for each patient, it computes the prediction M_1 would make if the patient had just been measured: $y_1^{\text{fair}} = M_1(x \mid \delta_t{=}0,\ m_t{=}0)$. It then applies a per-patient correction $\delta_i = \alpha_r(y_{1,i}^{\text{fair}} - y_{\text{current},i})$, where y_{current} is y_1 on round 0 and $y_2^{(r-1)}$ thereafter, and α_r controls correction strength. Training targets are $y_2^{\text{target}} = \text{clip}(y_{\text{current}} + \delta,\ 0,\ 1)$.

RM_1 evaluates y_2 each round via severity-adjusted disparity, which isolates the gap attributable to measurement bias: $(\bar{y}_{2,g=0} - \bar{S}_{g=0}) - (\bar{y}_{2,g=1} - \bar{S}_{g=1})$,. The fairness score is $1 -$ |severity-adjusted disparity|.

Training and deployment. Each of R rounds applies: (1) frozen M_1 to obtain y_1; (2) RM_2 to correct y_{current} and retrain M_2; (3) updated M_2 to produce y_2, audited by RM_1. At deployment, M_2 is composed with M_1 so the threshold policy operates on debiased scores $y_2 = M_2(M_1(x),\ \delta_t,\ m_t)$. Measurement-rate equalization emerges from corrected risk estimates rather than an explicit rule-based override. The full algorithm is provided in 1 in the appendix.

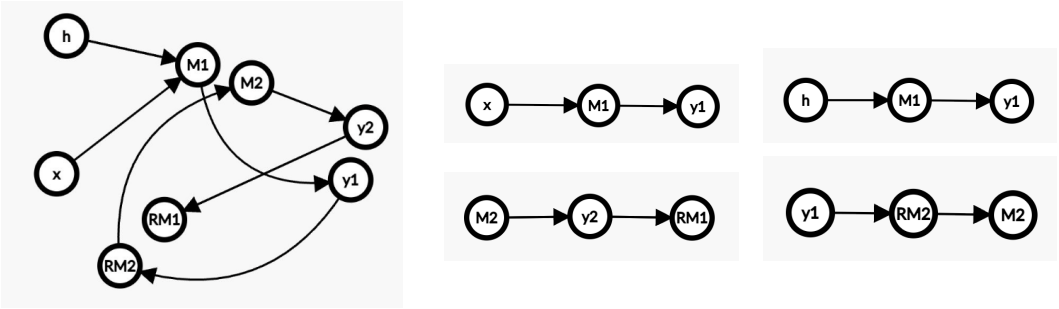

Figure 1: **Causal RLHF graph and its piecewise decomposition.** Left: the full causal graph with observed nodes split into temporal copies. Right: the three sequential causal chains. At $t{=}1$, $(x \to M_1 \to y_1)$ occurs. At $t{=}2$, $(M_2 \to y_2 \to \text{RM}_1)$ and $(h \to M_1 \to y_1)$ occur simultaneously. At $t{=}3$, $(y_1 \to \text{RM}_2 \to M_2)$ occurs. The unrolling breaks the feedback cycle into a valid DAG while the latent node h accounts for hidden confounders.

5 Experiments

5.1 Synthetic experiment

We simulate $N = 2000$ patients over $T = 20$ timesteps. Each patient i is assigned to a subgroup $g_i \in \{0, 1\}$ with equal prevalence ($N/2$ per group). The two groups have identical underlying health dynamics but differ systematically in their measurement rates, mimicking real-world disparities.

Severity state. Each patient has a hidden binary severity state $S_t \in \{0, 1\}$ evolving as a time-homogeneous Markov chain:

$$P(S_{t+1} = 1 \mid S_t = 1) = p_{11} = 0.8, \tag{5}$$
$$P(S_{t+1} = 1 \mid S_t = 0) = p_{01} = 0.2. \tag{6}$$

The stationary distribution gives $P(S = 1) = p_{01}/(1 - p_{11} + p_{01}) = 0.5$, so both groups have equal expected severity prevalence. Initial states are drawn from this stationary distribution.

Measurement process. At each timestep, a measurement decision $M_t \in \{0, 1\}$ is drawn from a Bernoulli whose parameter depends on both current severity and group membership:

$$P(M_t = 1 \mid S_t, g) = \begin{cases} 0.80 & \text{if } S_t = 1,\ g = 0 \\ 0.60 & \text{if } S_t = 1,\ g = 1 \\ 0.30 & \text{if } S_t = 0,\ g = 0 \\ 0.15 & \text{if } S_t = 0,\ g = 1 \end{cases} \tag{7}$$

This encodes two structural properties of real clinical data. First, sicker patients are measured more often regardless of group. Second, Group 1 is systematically under-measured across all severity levels, resulting in a measurement rate disparity of 17.7 percentage points (Group 0: 55.1%, Group 1: 37.4%) despite identical true severity distributions (Group 0: 50.7%, Group 1: 50.4%).

Observed feature. When $M_t = 1$, a biomarker value is drawn as:

$$X_t \sim \mathcal{N}(\mu_{S_t}, \sigma^2), \quad \mu_1 = 1.0,\ \mu_0 = 0.0,\ \sigma = 1.0. \tag{8}$$

When $M_t = 0$, X_t is missing (NaN), representing a noisy clinical variable whose distribution differs by severity, analogous to a lab value such as troponin or creatinine.

Outcome label. The outcome $Y_i = S_T^{(i)} \oplus \varepsilon_i$, $\varepsilon_i \sim \text{Bernoulli}(0.05)$ is the patients's last timestep severity, yielding approximately 50% prevalence in both groups. The noise term reflects the imperfect correspondence between clinical labels and true underlying severity in real EHR data.

Feature construction. From the raw trajectory data we construct four features per (patient, timestep) pair: (1) $\tilde{x}_t$, the forward-filled last observed value (global mean if never observed); (2) $m_t \in \{0, 1\}$, missingness indicator for the current timestep; (3) $\delta_t \in \mathbb{Z}_{\geq 0}$, the gap in timesteps since the last measurement; and (4) $c_t \in \mathbb{Z}_{\geq 0}$, the cumulative measurement count.

Bias mitigation setup. Both mitigation methods are evaluated on the synthetic dataset. We compare three conditions against the same deployment simulation described in Task 2: the biased baseline (M_1 under the standard threshold policy), the uncertainty-triggered policy (Task 3), and the Causal RLHF pipeline (Task 4). For the uncertainty-triggered policy, $u_0 = 0.20$ and $g_0 = 3$; for Causal RLHF, $R = 8$ rounds with annealing schedule $\alpha_r = \min(0.4 + 0.12r,\ 1.0)$ and blending rate $\eta = 0.6$. M_2 is instantiated as an ordinary least squares linear regression with standardised inputs. All hyperparameters were selected by manual tuning; no grid search was performed.

5.2 MIMIC-IV experiment

We apply the same bias identification pipeline to MIMIC-IV, a publicly available dataset of hospital admissions from Beth Israel Deaconess Medical Center [11, 10, 9], accessed under institutional ethics training and a signed data use agreement. Unlike the synthetic experiment, the true underlying health state is never observed in real clinical data — the MIMIC experiment instead asks whether the preconditions for the feedback loop are present in practice.

Cohort construction. We restrict to adults (`anchor_age` ≥ 18) with a first ICU stay only and minimum length of stay of 48 hours, ensuring sufficient temporal resolution for time-series feature construction. We further restrict to Private or Medicaid insurance, excluding Medicare, to avoid age-related confounds. This yields 12,726 patients: 8,301 Private (Group 0) and 4,425 Medicaid (Group 1). Private patients constitute the reference group and Medicaid patients the disadvantaged group, used as a proxy for access disparity rather than a causal claim.

Outcome variable. In-hospital mortality serves as Y ($Y = 1$ if `deathtime` is non-null, $Y = 0$ otherwise). Overall mortality is 10.58% (Private: 9.61%, Medicaid: 12.38%). The higher Medicaid mortality indicates genuine severity differences between groups, which we acknowledge as a limitation: unlike the synthetic experiment, equal underlying severity cannot be guaranteed.

Lab variable selection. We systematically evaluated 12 candidate variables from `labevents`. Routine electrolytes and metabolic panels showed no disparity or slight reverse disparity (Private/Medicaid ratios 0.94–0.96). Troponin T (itemid 51003) was the sole variable with higher measurement rates among Private patients (ratio 1.062), with a 6.6 percentage-point zero-measurement gap among deceased patients (Private: 57.1%, Medicaid: 63.7%) absent among survivors. Creatinine (itemid 50912) is retained as a control given its near-universal ordering rate (99.9% Private, 100% Medicaid).

Feature construction. Each ICU stay is divided into 12-hour bins up to a maximum of 20 timesteps. The bin size is motivated by the median inter-measurement interval for troponin (9.42 hours). We construct the same four features — $\tilde{x}_t$, m_t, δ_t, c_t — separately for troponin and creatinine, yielding eight features per (patient, timestep) pair, with Y and group label joined onto every row. The final feature matrix contains 123,342 rows across 12,726 patients (mean 9.69 timesteps per patient).

6 Results

6.1 Bias identification

6.1.1 Synthetic experiment

Models trained on biased measurement data learn to treat unmeasuredness as a proxy for health, and this observability shortcut is statistically dominant over genuine clinical signal. The fitted coefficients directly reveal the shortcut (Table 1): both observability features carry highly significant negative weights (m_t: -0.130, $p < 0.001$; δ_t: -0.035, $p < 0.001$), while the clinical signal $\tilde{x}_t$ is only weakly significant ($p = 0.015$) with a substantially smaller effect size, and measurement count is non-significant ($p = 0.31$). The model relies more on the measurement pattern than its content, producing a lower mean predicted risk for Group 1 (0.507) than for Group 0 (0.521) despite equal severity.

Table 1: **Logistic regression coefficients with Wald test statistics.** The model relies more heavily on measurement-pattern features (m_t, δ_t) than on the clinical signal ($\tilde{x}_t$), confirming the observability shortcut. Asterisks denote significance level: $***p < 0.001$, $*p < 0.05$, ns = not significant.

Feature	Coef	z	p-value	Sig
$\tilde{x}_t$ (forward-filled)	$+0.026$	2.42	0.015	*
m_t (missingness indicator)	-0.130	-4.42	< 0.001	***
δ_t (gap length)	-0.035	-4.73	< 0.001	***
c_t (measurement count)	$+0.004$	1.02	0.309	ns

The observability shortcut manifests in individual predictions: masking a measurement reduces predicted risk regardless of true clinical state. The counterfactual masking audit confirms the shortcut operates on real predictions: mean Δrisk is positive for both groups (Group 0: $+0.026$; Group 1: $+0.020$). The smaller effect in Group 1 showed that the model had so thoroughly absorbed their unmeasuredness into its baseline predictions that masking an additional step barely changed anything, providing evidence of the shortcut's depth. The multi-step masking audit further shows that predicted risk degrades monotonically with k, and the inter-group disparity persists through $k = 10$ (Table 2, Figure B5), driven by the staleness of Group 1's forward-filled values.

The observability shortcut is not an artifact of model simplicity; it emerges across all tested model classes through different but equivalent pathways. We additionally trained an MLP (64 units, ReLU) and XGBoost on identical features and splits. All three models achieve near-identical AUROC scores (LR: 0.530, MLP: 0.538, XGB: 0.518), confirming task difficulty is model-independent. SHAP-based feature attribution reveals all three exploit the observability shortcut through different pathways (Table B5): logistic regression through m_t and δ_t directly, the MLP through δ_t, and XGBoost through $\tilde{x}_t$, which is itself a product of the measurement process as longer gaps produce increasingly stale forward-filled values. The shortcut is not eliminated by model sophistication; it is redistributed across features.

Once encoded in a deployed model, the observability shortcut creates a self-sustaining feedback loop that compounds measurement disparity without any new external bias. The policy simulation reveals three coupled disparity channels sustaining simultaneously across 30 deployment timesteps (Table 3, Figure B6). Under the baseline policy, Group 0 is measured at a 4.6 percentage point higher rate than Group 1 despite equal true severity, with mean predicted risk remaining 1.1 percentage points higher and gaps 0.576 timesteps shorter throughout deployment. The policy perpetuates lower measurement rates for Group 1, keeping their features staler and their predicted risk lower in a self-reinforcing cycle that compounds rather than corrects the historical deficit. Critically, true severity remains equal throughout, confirming none of the observed disparity reflects genuine clinical differences. The feedback loop requires no new external bias to sustain itself.

Table 2: **Predicted risk under multi-step masking.** Predicted risk decays monotonically as consecutive unmeasured steps increase, and the inter-group disparity persists through $k = 10$, showing the shortcut compounds with prolonged unobservation.

k	Group 0	Group 1	G0−G1
0	0.526	0.508	0.018
3	0.482	0.470	0.012
5	0.465	0.453	0.012
10	0.422	0.410	0.012

Table 3: **Disparity (Group 0 − Group 1) under baseline policy.** The feedback loop sustains measurement, risk, and gap disparities throughout deployment despite equal true severity. Positive values indicate Group 1 disadvantage for measurement rate and predicted risk; negative values indicate Group 1 disadvantage for gap length.

Metric	Baseline disparity
Measurement rate	$+0.046$
Predicted risk	$+0.011$
Gap length	-0.576

6.1.2 MIMIC-IV experiment

The observability shortcut replicates in real clinical data, operating more strongly for selectively ordered labs than for routine controls. The coefficient on `trop_missing_now` is -0.2794 ($p < 0.001$, ***), confirming that being unmeasured for troponin predicts lower mortality risk (Table 4). The effect is larger for troponin than creatinine (-0.28 vs -0.16), consistent with troponin's more selective ordering carrying a stronger informative missingness signal. The troponin gap coefficient is positive and significant ($p = 0.013$), reflecting a clinically coherent reversal: patients who receive troponin after a long gap are those who received enough clinician concern to trigger the order. The counterfactual masking audit produces smaller Δrisk values than the synthetic experiment (Private: $+0.00112$; Medicaid: $+0.00081$), which we attribute to troponin being missing in 95.5% of bins. The model has absorbed unmeasuredness so thoroughly that the additional masking step has little impact.

The gap disparity signature persists in real deployment simulation, though predicted risk and measurement rate disparities do not replicate cleanly. Medicaid patients enter with longer troponin gaps (9.04 vs 8.53 bins), and this disparity persists throughout the 30-step simulation (5.45 vs 5.34 bins at the final timestep). However, predicted risk and measurement rate show no consistent directional disparity: predicted risks cluster around 0.13, well below $\tau = 0.5$, collapsing the policy to a near-uniform low-measurement regime where group differences cannot manifest. The gap disparity persists because Medicaid patients initialise with longer gaps and the policy is insufficiently sensitive to correct them. These results might understate real-world risk, as the model is frozen during simulation whereas real deployments involve periodic retraining that could reinforce the bias.

6.2 Bias mitigation

Our bias mitigation techniques were effective at reducing bias from the baseline model. Figure 2 compares both methods against the baseline across three disparity metrics. For measurement rate (left column), both methods yield identical results, reducing the baseline disparity of +0.046 to -0.005. This shows a near-complete correction that slightly over-corrects, with Group 1 ending up marginally more measured than Group 0. For predicted risk (center column), both methods substantially reduce the baseline disparity of +0.011: uncertainty-triggered measurement corrects within five timesteps while causal RLHF eliminates the gap almost immediately, achieving average disparities of +0.0034 and +0.0030, respectively. Causal RLHF in this area performs marginally better. Some residual disparity remains, indicating the bias is not fully eliminated. For gap length (right column), both methods collapse the baseline disparity of -0.576 to near zero ($-$ 0.0002), indicating that both groups are re-measured at comparable frequencies.

Table 4: **Logistic regression coefficients for the MIMIC-IV model comparing troponin and creatinine features.** The observability shortcut is stronger for selectively ordered troponin, confirming the bias scales with measurement selectivity. Asterisks: $***p < 0.001$, $*p < 0.05$, ns = not significant.

Feature	Coef	p-value	Sig
`trop_missing_now`	-0.2794	< 0.001	***
`trop_gap`	$+0.0064$	0.013	*
`creat_missing_now`	-0.1595	< 0.001	***
`creat_gap`	-0.0229	0.174	ns

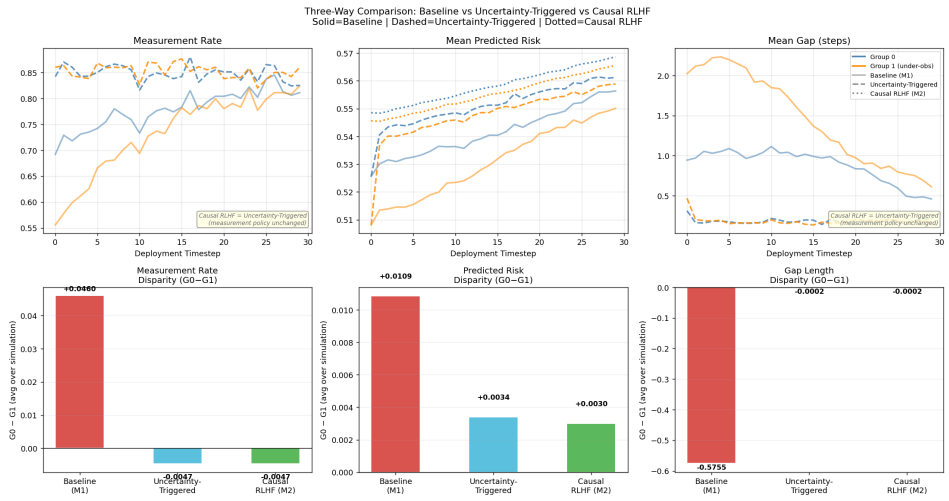

Figure 2: **Baseline (solid) vs. uncertainty-triggered mitigation (dashed) vs. causal RLHF (dotted)** Both mitigation methods nearly eliminate disparity across all three metrics, with causal RLHF achieving this through corrected risk scores alone rather than policy overrides.

7 Conclusions, implications, and limitations

Our bias identification methods confirmed that models trained on biased measurement data learn to treat a lack of data as a proxy for health. Furthermore, deploying these models under a risk-driven monitoring policy creates a self-sustaining feedback loop that compounds measurement disparity without any new external bias. Our bias mitigation methods were relatively effective at reducing bias from the baseline model, though they did not eliminate it entirely.

We agree that informative missingness improves prediction in clinical settings and helps overcome past barriers associated with missing data. However, our research shows that the predictive signal of informative missingness becomes a mechanism for compounding inequality and imbalance in deployment. Despite this, it is important to note that highly beneficial techniques in clinical settings, such as human validation and iterative improvements to downstream training, require a feedback loop. Therefore, our research also provides bias-mitigation strategies that preserve the benefits of feedback loops in clinical settings while addressing potential consequences of performative prediction.

Our findings are subject to several limitations. The synthetic experiment utilized clean, controlled bias injection, but real-world measurement disparities are noisier. Our policy simulation is also retrospective, as it shows what would happen under the programmed policy, not what occurs in real clinical deployment. In our MIMIC analysis, the true underlying severity is never directly observed, so the troponin disparity may reflect genuine clinical differences between insurance groups. Furthermore, the effect sizes on MIMIC were small, since ICU protocols partially equalize measurement intensity and limit how strongly a feedback loop can operate in this setting. Regarding our bias mitigation techniques, our primary limitation was that the causal assumptions we introduced created a controlled setting that may not generalize to complex clinical settings. Future work will focus on resolving this limitation by better articulating which components influence a healthcare setting and how they can be represented and manipulated in a lab-controlled environment.

References

[1] G. A. Adam, C.-H. K. Chang, B. Haibe-Kains, and A. Goldenberg. Hidden risks of machine learning applied to healthcare: Unintended feedback loops between models and future data causing model degradation. In F. Doshi-Velez, J. Fackler, K. Jung, D. Kale, R. Ranganath, B. Wallace, and J. Wiens, editors, *Proceedings of the 5th Machine Learning for Healthcare Conference*, volume 126 of *Proceedings of Machine Learning Research*, pages 710–731. PMLR, 07–08 Aug 2020. URL https://proceedings.mlr.press/v126/adam20a.html.

[2] N. C. Arpey, A. H. Gaglioti, and M. E. Rosenbaum. How socioeconomic status affects patient perceptions of health care: A qualitative study. *Journal of Primary Care & Community Health*, 8(3):169–175, 2017. doi: 10.1177/2150131917697439. URL https://doi.org/10.1177/2150131917697439. PMID: 28606031.

[3] Z. Che, S. Purushotham, K. Cho, D. Sontag, and Y. Liu. Recurrent neural networks for multivariate time series with missing values, 2016. URL https://arxiv.org/abs/1606.01865.

[4] F. Cismondi, A. S. Fialho, S. M. Vieira, S. R. Reti, J. M. Sousa, and S. N. Finkelstein. Missing data in medical databases: Impute, delete or classify? *Artificial Intelligence in Medicine*, 58 (1):63–72, 2013. ISSN 0933-3657. doi: https://doi.org/10.1016/j.artmed.2013.01.003. URL https://www.sciencedirect.com/science/article/pii/S0933365713000055.

[5] J. Cross, M. Choma, and J. Onofrey. Bias in medical ai: Implications for clinical decision-making. *PLOS Digital Health*, 3, 11 2024. doi: 10.1371/journal.pdig.0000651.

[6] A. R. Dalton, A. Bottle, M. Soljak, C. Okoro, A. Majeed, and C. Millett. The comparison of cardiovascular risk scores using two methods of substituting missing risk factor data in patient medical records. *Informatics in primary care*, 19(4), 2011.

[7] M. De-Arteaga, A. Dubrawski, and A. Chouldechova. Learning under selective labels in the presence of expert consistency, 2018. URL https://arxiv.org/abs/1807.00905.

[8] D. Ensign, S. A. Friedler, S. Neville, C. Scheidegger, and S. Venkatasubramanian. Runaway feedback loops in predictive policing. In *Conference on fairness, accountability and transparency*, pages 160–171. PMLR, 2018.

[9] A. Goldberger, L. Amaral, L. Glass, J. Hausdorff, P. C. Ivanov, R. Mark, and H. E. Stanley. Physiobank, physiotoolkit, and physionet: Components of a new research resource for complex physiologic signals. *Circulation*, 101(23):e215–e220, 2000. RRID:SCR_007345.

[10] A. Johnson, L. Bulgarelli, T. Pollard, B. Gow, B. Moody, S. Horng, L. A. Celi, and R. Mark. MIMIC-IV (version 3.1), 2024. RRID:SCR_007345.

[11] A. E. W. Johnson, L. Bulgarelli, L. Shen, et al. MIMIC-IV, a freely accessible electronic health record dataset. *Scientific Data*, 10(1), 2023. doi: 10.1038/s41597-022-01899-x.

[12] H. Lakkaraju, J. Kleinberg, J. Leskovec, J. Ludwig, and S. Mullainathan. The selective labels problem: Evaluating algorithmic predictions in the presence of unobservables. In *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 275–284, 2017.

[13] N. Pagan, J. Baumann, E. Elokda, G. De Pasquale, S. Bolognani, and A. Hannák. A classification of feedback loops and their relation to biases in automated decision-making systems. In *Equity and Access in Algorithms, Mechanisms, and Optimization*, EAAMO '23, page 1–14. ACM, Oct. 2023. doi: 10.1145/3617694.3623227. URL http://dx.doi.org/10.1145/3617694.3623227.

[14] R. B. Parikh, S. Teeple, and A. S. Navathe. Addressing bias in artificial intelligence in health care. *JAMA*, 322(24):2377–2378, 12 2019. ISSN 0098-7484. doi: 10.1001/jama.2019.18058. URL https://doi.org/10.1001/jama.2019.18058.

[15] J. C. Perdomo, T. Zrnic, C. Mendler-Dünner, and M. Hardt. Performative prediction, 2021. URL https://arxiv.org/abs/2002.06673.

[16] S. R. Pfohl, A. Foryciarz, and N. H. Shah. An empirical characterization of fair machine learning for clinical risk prediction. *Journal of Biomedical Informatics*, 113:103621, 2021. ISSN 1532-0464. doi: https://doi.org/10.1016/j.jbi.2020.103621. URL https://www.sciencedirect.com/science/article/pii/S1532046420302495.

[17] M. Phelan, N. Bhavsar, and B. Goldstein. Illustrating informed presence bias in electronic health records data: How patient interactions with a health system can impact inference. *eGEMs (Generating Evidence Methods to improve patient outcomes)*, 5:22, 12 2017. doi: 10.5334/egems.243.

[18] R. Rios, R. J. Miller, N. Manral, T. Sharir, A. J. Einstein, M. B. Fish, T. D. Ruddy, P. A. Kaufmann, A. J. Sinusas, E. J. Miller, T. M. Bateman, S. Dorbala, M. Di Carli, S. D. Van Kriekinge, P. B. Kavanagh, T. Parekh, J. X. Liang, D. Dey, D. S. Berman, and P. J. Slomka. Handling missing values in machine learning to predict patient-specific risk of adverse cardiac events: Insights from refine spect registry. *Computers in Biology and Medicine*, 145:105449, 2022. ISSN 0010-4825. doi: https://doi.org/10.1016/j.compbiomed.2022.105449. URL https://www.sciencedirect.com/science/article/pii/S0010482522002414.

[19] R. Sisk, L. Lin, M. Sperrin, J. K. Barrett, B. Tom, K. Diaz-Ordaz, N. Peek, and G. P. Martin. Informative presence and observation in routine health data: A review of methodology for clinical risk prediction. *Journal of the American Medical Informatics Association*, 28(1):155–166, 11 2020. ISSN 1527-974X. doi: 10.1093/jamia/ocaa242. URL https://doi.org/10.1093/jamia/ocaa242.

[20] M. Sun, M. Engelhard, A. Bedoya, and B. Goldstein. Incorporating informatively collected laboratory data from ehr in clinical prediction models. *BMC Medical Informatics and Decision Making*, 24, 07 2024. doi: 10.1186/s12911-024-02612-1.

[21] A. L. Tan, E. J. Getzen, M. R. Hutch, Z. H. Strasser, A. Gutiérrez-Sacristán, T. T. Le, A. Dagliati, M. Morris, D. A. Hanauer, B. Moal, C.-L. Bonzel, W. Yuan, L. Chiudinelli, P. Das, H. G. Zhang, B. J. Aronow, P. Avillach, G. Brat, T. Cai, C. Hong, W. G. La Cava, H. Hooi Will Loh, Y. Luo, S. N. Murphy, K. Yuan Hgiam, G. S. Omenn, L. P. Patel, M. Jebathilagam Samayamuthu, E. R. Shriver, Z. Shakeri Hossein Abad, B. W. Tan, S. Visweswaran, X. Wang, G. M. Weber, Z. Xia, B. Verdy, Q. Long, D. L. Mowery, and J. H. Holmes. Informative missingness: What can we learn from patterns in missing laboratory data in the electronic health record? *Journal of Biomedical Informatics*, 139:104306, 2023. ISSN 1532-0464. doi: https://doi.org/10.1016/j.jbi.2023.104306. URL https://www.sciencedirect.com/science/article/pii/S1532046423000278.

A Causal RLHF earlier progressions

These illustrate the causal RLHF graphs we originally constructed, which helped us identify violations of the DAG assumption and observability conflicts.

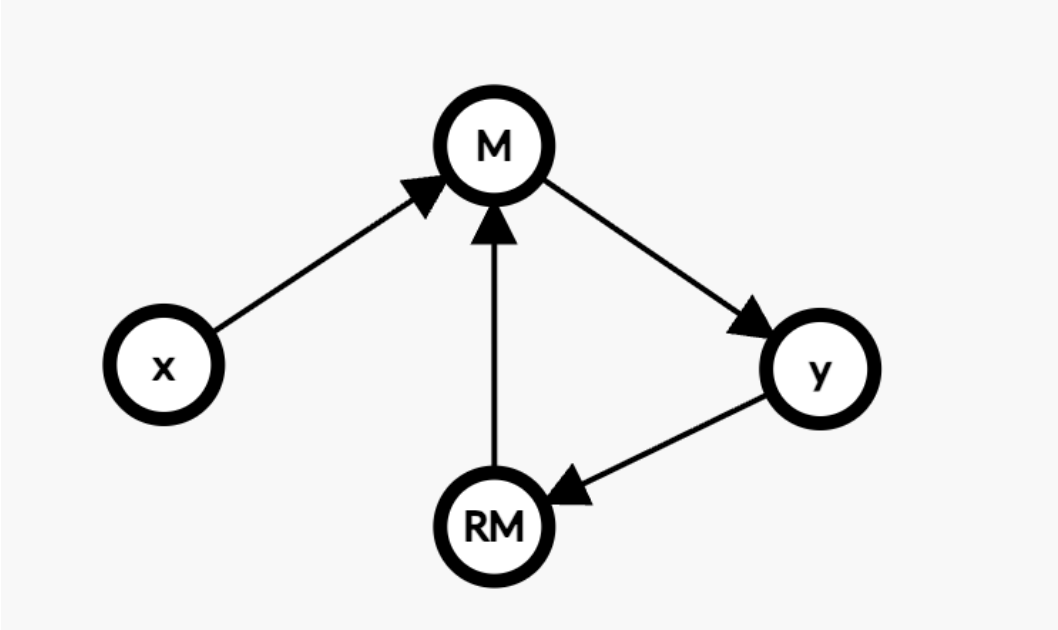

Figure A3: **Original causal graph of the RLHF pipeline.** Let x be the user input, M be the ML model that receives the user input, y be the generated result, and RM be the reward model. Notice how this version of the causal RLHF pipeline creates a cycle and does not account for hidden confounders, introducing two challenges that we address in our preliminaries section.

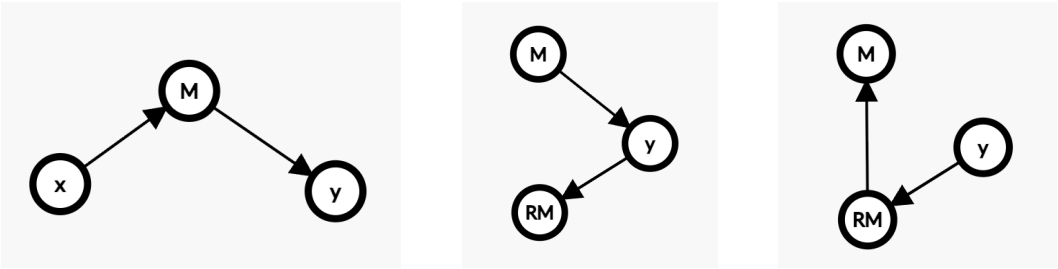

Figure A4: **Piecewise separation of the causal RLHF graph.** We separate the original causal representation of RLHF into three piecewise causal graphs. This makes the sequence into a series of DAGs. It also sets up the problem for applying d-separation to determine conditional independence within each piecewise network. However, doing so revealed observability problems regarding nodes having to go from being observed to being unobserved between timesteps. This challenge was addressed in our preliminaries section.

B Additional figures related to bias identification task results

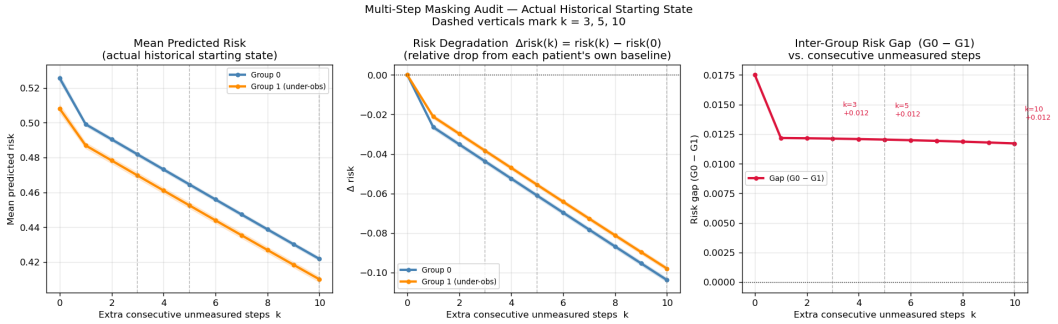

Figure B5: **Multi-step masking audit.** Left: predicted risk vs. consecutive unmeasured steps k by group. Centre: risk degradation relative to k=0 baseline. Right: inter-group disparity vs. k. Risk declines monotonically for both groups, confirming the model systematically underestimates risk for unobserved patients.

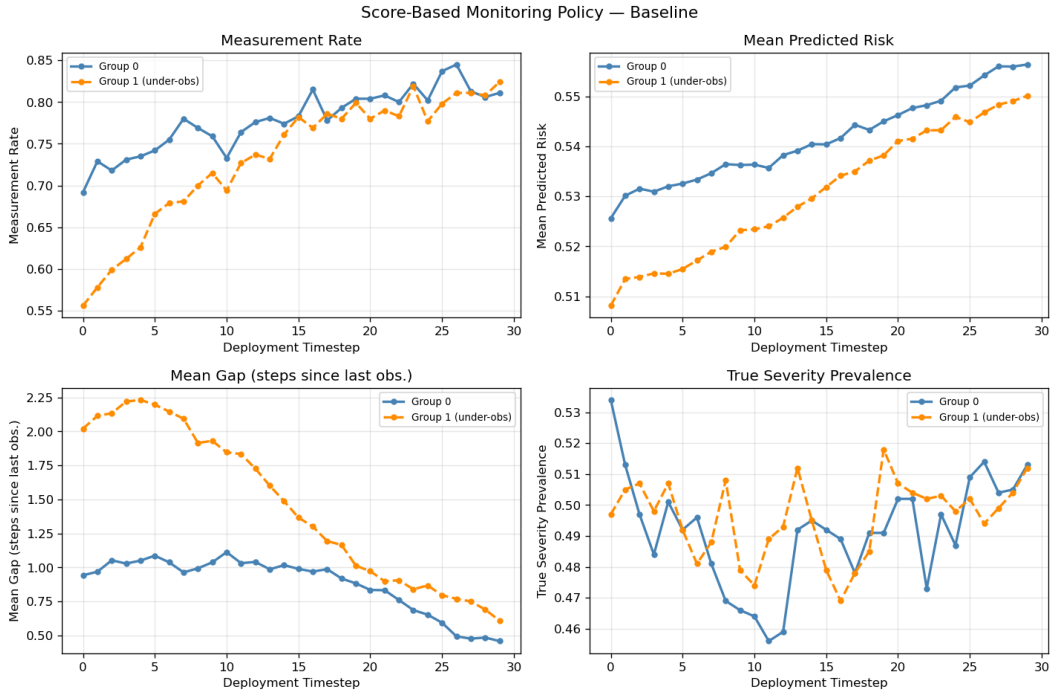

Figure B6: **Baseline policy simulation over 30 deployment timesteps.** Measurement rate (top left), predicted risk (top right), gap length (bottom left), and true severity (bottom right) by group. The feedback loop sustains all three disparity channels while true severity remains equal, confirming the disparity is model-induced, not clinical.

Table B5: **Feature importance across model classes.** All three models exploit the observability shortcut through different pathways: LR via m_t and δ_t directly, MLP via δ_t, XGBoost via stale forward-filled values. The shortcut is redistributed across features instead of being eliminated by model sophistication.

Feature	LR coef	LR rank	MLP |SHAP|	MLP rank	XGB |SHAP|	XGB rank
$\tilde{x}_t$ (forward-filled)	$+0.0240$	3	0.0112	4	0.2837	1
m_t (missingness indicator)	-0.1054	1	0.0117	3	0.0715	4
δ_t (gap length)	-0.0386	2	0.0352	1	0.0882	3
c_t (measurement count)	$+0.0058$	4	0.0133	2	0.1801	2

C Causal RLHF algorithm

Algorithm 1 Causal RLHF Pipeline

Require: Frozen primary model M_1, feature matrix $X = [\tilde{x}_t,\ m_t,\ \delta_t,\ c_t]$, true severity S, group labels g (diagnostics only), number of rounds R, blending rate η, annealing parameters $\alpha_0 = 0.4$, $\Delta\alpha = 0.12$

Initialise M_2 (unfitted), RM_1, RM_2
$y_1 \leftarrow M_1$.predict_proba(X)
for $r = 0$ **to** $R - 1$ **do**
  // Timestep t=1: $x \to M_1 \to y_1$
  $y_1 \leftarrow M_1$.predict_proba(X) ▷ M_1 observed, frozen
  if M_2 is fitted **then**
    $y_{\text{current}} \leftarrow \text{clip}\big(M_2.\text{predict}(y_1, \delta_t, m_t),\ 0,\ 1\big)$
  else
    $y_{\text{current}} \leftarrow y_1$
  end if

  // Timestep t=3: $y_1 \to$ RM$_2$ $\to M_2$
  $X^{\text{fair}} \leftarrow X$ with $\delta_t \leftarrow 0,\ m_t \leftarrow 0$ ▷ Counterfactual features
  $y_1^{\text{fair}} \leftarrow M_1$.predict_proba$(X^{\text{fair}})$
  $\alpha_r \leftarrow \min(\alpha_0 + \Delta\alpha \cdot r,\ \ 1.0)$ ▷ Annealing schedule
  $\boldsymbol{\delta} \leftarrow \alpha_r \cdot (y_1^{\text{fair}} - y_{\text{current}})$ ▷ Residual correction
  $y_2^{\text{target}} \leftarrow \text{clip}(y_{\text{current}} + \boldsymbol{\delta},\ \ 0,\ \ 1)$
  Define $Z \leftarrow [y_1,\ \ \delta_t,\ \ m_t,\ \ y_1 \cdot \delta_t,\ \ y_1 \cdot m_t]$
  $Z_s \leftarrow \text{standardise}(Z)$
  Fit M_2^{new} via OLS on $(Z_s,\ y_2^{\text{target}})$
  if M_2 was previously fitted **then**
    $y_2^{\text{blend}} \leftarrow (1 - \eta)\ \text{clip}(M_2(Z_s), 0, 1) + \eta\ \text{clip}(M_2^{\text{new}}(Z_s), 0, 1)$
    Refit M_2 via OLS on $(Z_s,\ y_2^{\text{blend}})$ ▷ Soft update
  else
    $M_2 \leftarrow M_2^{\text{new}}$
  end if

  // Timestep t=2: $M_2 \to y_2 \to$ RM$_1$; $h \to M_1 \to y_1$
  $y_2 \leftarrow \text{clip}\big(M_2.\text{predict}(y_1, \delta_t, m_t),\ 0,\ 1\big)$ ▷ Debiased predictions
  $d_{\text{adj}} \leftarrow (\bar{y}_{2,g=0} - \bar{S}_{g=0}) - (\bar{y}_{2,g=1} - \bar{S}_{g=1})$ ▷ RM_1 fairness audit
  $\text{score}_r \leftarrow 1 - |d_{\text{adj}}|$
end for

return M_2 composed with M_1
▷ Deployment: $\hat{y} = \text{clip}\big(M_2(M_1(x),\ \delta_t,\ m_t),\ 0,\ 1\big)$