# ZERO-SUM BIAS CORRECTIONS FOR CONFOUNDING BY INDICATION IN OFFLINE REINFORCEMENT LEARNING

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Although offline reinforcement learning promises more personalized treatment regimes in clinical care, challenges remain when action- and outcome-relevant information goes unrecorded in the data used to train policies. Standard value-based offline reinforcement learning methods converge to the observational value of actions and conflate the reason the action was chosen with its effect. Clinical data commonly suffers from confounding by indication, where patients with unrecorded indicators of disease severity both receive more aggressive treatment and experience worse outcomes. This biases naive learners, undervaluing life-saving treatments while overvaluing inaction. Many confounding-robust offline reinforcement learning methods impose bounds that further penalize treatment, while others require strict identification conditions that may be false in practice. In both cases, the learned policy can collapse and fail to provide any patient-specific targeting at all. We introduce an offline reinforcement learning algorithm to correct the training targets under static confounders. Our correction is derived from an omitted-variable-bias factorization and approximates the bias by the product of two components estimated from the full trajectory history, with the direction of the correction given by the data as the sign of a within-trajectory covariance. As per-action biases at a state provably cancel, we constrain our correction to be zero-sum across actions so that value is redistributed rather than uniformly suppressed. Across three synthetic medical environments, one fit to real clinical dynamics, our method produces higher-value policies than competing deconfounding and offline reinforcement learning approaches. Where conservative baselines collapse, our method continues to differentiate between patients.

## 1 INTRODUCTION

Recent advances in reinforcement learning (RL) have shown strong performance in domains such as robotics and natural language processing, and have attracted a sustained interest in healthcare for dynamic treatment regimes (Yu et al., 2021). However fundamental challenges remain, in particular with regard to learning policies from static, historical data (*Offline* RL) rather than in online settings (Levine et al., 2020). Oftentimes, the historical data is generated by behavior policies using information beyond that which is contained in the logged data (Zhao et al., 2020).

In offline RL, confounding occurs when the behavior policy that generates the logged dataset takes actions based on the value of a variable that also contributes to the outcome. This becomes unobserved confounding when the variable is not recorded in the logged dataset (Gottesman et al., 2019). Observed confounding can be mitigated by standard adjustment (Pearl, 2010), and many methods have been developed to perform adjustment to learn policies in sequential decision-making settings (Bica et al., 2020; Melnychuk et al., 2022; Namkoong et al., 2020). However, unobserved confounding is significantly more challenging to mitigate, and cannot be point-identified without assumptions on the underlying structural causal model (SCM) governing the data (Pearl, 2001).

Confounding-by-indication (CBI) occurs when the confounder drives both the expected outcome and the probability of certain actions being taken. In this work, we focus on a type of confounding-by-indication, confounding-by-severity (CBS), where a more 'severe' confounder status leads to



Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

more intensive treatment and worse outcomes (Kyriacou & Lewis, 2016). In clinical settings with CBS, patients who receive a greater number of treatments have worse outcomes. For example, patients with severe sepsis are likely to receive much more treatment compared to healthy patients yet still die at a higher rate (Brand et al., 2017). If the patient's sepsis status is not recorded in the logged data, vasopressors may appear *harmful*.

Many offline RL methods apply pessimism to address different sources of uncertainty. These methods work by reducing the value of actions in states for which the actions are rare or for which the observed dynamics may have been produced by many different generative processes (Kumar et al., 2020; Pace et al., 2024). We show that under sequential ignorability given the full state, positivity, and no aggregate selection on gains, the per-action biases sum to zero at each observation. As a result, any non-negative penalty drives the propensity-weighted mean error strictly negative, and at every observation there is an action the logged data already under-values. A penalty that is positive on that action estimates it strictly less accurately than making no adjustment at all.

Our contributions are as follows:

1. We prove that under a set of realistic assumptions, the per-action biases at a state are zero-sum. As a consequence, pessimistic methods that subtract from already underestimated action values increase the value estimation error.
2. We show that under a static confounder, both factors of the omitted-variable-bias decomposition of Chernozhukov et al. (2026) are estimable from the panel structure of the logged data alone.
3. We introduce a new signed correction method (ZSOC) for learning confounding-robust offline RL policies in settings with CBS. We evaluate ZSOC on our benchmark against offline RL and deconfounding baselines, and show that it leads to higher policy returns on the tested environments.

## 2 Related Work

**Off Policy Estimation (OPE).** OPE estimates the value of a fixed target policy from logged data, and a substantial literature addresses unobserved confounding in that setting (Deng et al., 2023; Zeng et al., 2025). Though confounding biases the OPE estimand in the same way as in our task, OPE is directed towards evaluating a given policy rather than learning a new policy.

**Instrument Variables and Proximal Reinforcement Learning.** Previous works have used instrument variables (Wu et al., 2025; Liao et al., 2024) or proximal causal inference (Tchetgen Tchetgen et al., 2020; 2024) to identify causal quantities. These methods require strong assumptions about the causal relationships between variables, but are able to point-identify causal effects when the assumptions hold. Bennett & Kallus (2024) introduced the proximal RL framework for OPE, while Lu et al. (2023) introduced a policy learning method based on proximal RL to learn policies under confounded POMDPs.

**Pessimistic Corrections.** Pessimistic methods underweight the observed action values based on some measure of confidence. Conservative Q-learning (CQL) (Kumar et al., 2020) penalizes actions that are out of distribution to obtain conservative Q-value estimates. CQL is intended to address distribution shift and not directly address confounding. The Marginal Sensitivity Model (MSM) based approaches build on the work of Rosenbaum (2002) and Tan (2006). These methods assume that the ratio between the observed propensity and the confounded propensity is bounded by a parameter $\Lambda$. For binary actions, the MSM model assumes $\frac{1}{\Lambda} \leq \frac{\pi_b(1|o,S)/(1-\pi_b(1|o,S))}{\pi_b(1|o)/(1-\pi_b(1|o))} \leq \Lambda$. Bruns-Smith & Zhou (2023) extend the MSM to the sequential setting, replacing the Bellman target with the worst-case pseudo-outcome inside the $\Lambda$ odds-ratio set at each decision point. However these methods require the $\Lambda$ to be specified by the user. Under-estimating $\Lambda$ can leave residual confounding uncorrected and the learned policy remains biased, while over-estimating $\Lambda$ can lead to policies that collapse towards always taking the action that is safest under the worst case.

**Latent Variable Estimation.** Many latent variable approaches fit models of treatment assignment and transitions using the logged data. These methods aim to directly estimate treatment effects and include the CEVAE (Louizos et al., 2017), Counterfactual Recurrent Networks (Bica et al.,

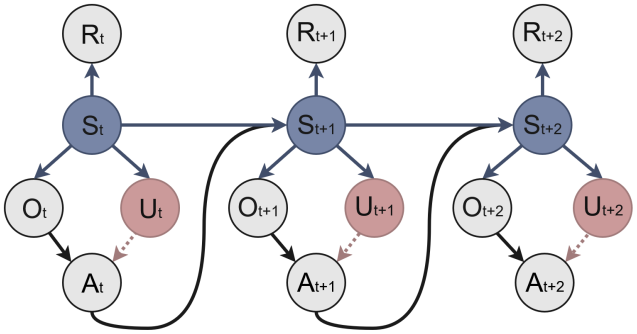



Figure 1: The SCM assumed for our tasks for the behavior policy. The state $S$ (blue) is unobserved by both the behavior policy and the learner and is a function of the previous state and the action only. Each state emits a noisy read $U$ (red) of the static confounder channel $\mathrm{sev}$, visible only to the behavior policy, and a set $\{O, A, R\}$ (grey) visible to both the behavior policy and the learner. The behavior policy uses $(O, U)$ to determine the action $A$ at each timepoint. Arrows indicate the causal relationships.

2020), Sequential Deconfounders (Hatt & Feuerriegel, 2024), and Causal Transformers (Melnychuk et al., 2022). Other latent variable approaches such as Deconfounded Actor-Critic (Lu et al., 2018) estimate a representation of the confounder given the trajectory history and condition the learner on the representation at test time. Though a confounder may be easy to recover during training because the full trajectory is available, when little history exists such as near the beginning of a new trajectory, recovery may be poor.

In contrast to the above methods, ours estimates a signed correction during training and deploys an observation-only policy. We do not assume the presence of valid instrument or proxy variables, and do not require manual specification of confounding bounds.

## 3 Method

### 3.1 Problem Setup

In this work we consider Confounded POMDPs (Shi et al., 2022) $\{\mathcal{S}, \mathcal{O}, \mathcal{U}, \mathcal{A}, \mathcal{R}, \mathcal{T}\}$ following the causal structure given in Figure 1. $\mathcal{S}$ is the true state, $\mathcal{O}$ is the space of observable features, $\mathcal{U}$ is the space of confounders, $\mathcal{A}$ is the space of actions, $\mathcal{R}$ is the space of rewards, and $\mathcal{T}$ is the transition function.

We assume an emission model and static confounders, where both $O_t \sim P_O(\cdot|S_t)$ and $U_t \sim P_U(\cdot|S_t)$ are noisy emissions of the true state $S_t \in \mathbb{R}^k$. In order for unobserved confounding to matter, it must be that $O_t$ only partially reveals $\mathrm{sev}$. Otherwise, policies could account for the confounder by adjusting for $O_t$ alone. Specifically, $U_t$ is a noisy read of one static coordinate of the state, the severity channel $\mathrm{sev}$, drawn once per trajectory and never updated: $U_t = \mathrm{sev} + \epsilon_{U,t}$. The behavior policy samples from $\pi_b(\cdot|O_t, U_t)$ while $U_t$ is not available to the learner policy $\pi(a|o)$. The reward is a function of the state alone: taking $A_t$ in $S_t$ produces $S_{t+1}$ and the reward $R_{t+1} \sim P_R(\cdot|S_{t+1})$. No reward depends on $A_t$, $O_t$, or $U_t$ except through the state it produced.

$Q^o(o, a)$ is the observational Q-value of the action $a$ at the given observation $o$ which converges to $\mathbb{E}_{\mathcal{D}}[G|o, a]$, or the expected return $G_t = \sum_{k=t+1}^{T} \gamma^{k-t-1} R_k$ given $(o, a)$ under the logged dataset $\mathcal{D}$. In contrast, $Q^c(o, a)$ is the causal value of the action $a$ at the given observation $o$: $Q^c(o, a) = \mathbb{E}_{S \sim P(\cdot|o)}[\mathbb{E}[G|o, a, S]] = \mathbb{E}[m(o, a, S)|o]$. The bias is defined as the gap between the value we see $Q^o(o, a)$ and the value we want $Q^c(o, a)$: $\mathrm{bias}(o, a) = Q^o(o, a) - Q^c(o, a)$.

Let $m(o, a, S) = \mathbb{E}[G|o, S, \mathrm{do}(A = a)]$, or the expected return assigned to the state $S$ with emission $o$ if the intervention were to force action $a$ and then continue under the behavior policy $\pi_b(\cdot|O, U)$. Notably, $m(o, a, S)$ evaluates the intervention under the behavior policy and not the learner policy.

### 3.2 ASSUMPTIONS

**Assumption 1** (Sequential ignorability given the state $S$)**.** *The SCM is Markov with respect to $S$, so that $\mathbb{E}[G_t|S_t] = \mathbb{E}[G_t|S_t, S_{1:t-1}]$, and the action carries no information about the return beyond $(O_t, S_t)$: $\mathbb{E}[G|o, a, S] = m(o, a, S)$. Equivalently, conditioning on the full state, observing $A_t = a$ and intervening to set $A_t = a$ agree in expectation.*

**Assumption 2** (Positivity)**.** *Let $\pi_b(a|o, S) = \mathbb{E}[\pi_b(a|o, U)|S]$. We assume that $P(a|o) > 0 \ \ \forall a \in \mathcal{A}$, where $P(a|o) = \mathbb{E}[\pi_b(a|o, S)|o]$.*

**Assumption 3** (No effect modification by hidden variables)**.** *We assume the hidden state does not modify the treatment effect. The treatment effect is defined as $\tau_a(o, S) = m(o, a, S) - m(o, a_{ref}, S)$ for the effect of $a$ against a reference action $a_{ref}$ (no treatment). We assume that this does not depend on the hidden state, $\tau_a(o, S) = \tau_a(o)$.*

**Assumption 4** (No selection on gains)**.** *On average, the behavior policy picks actions depending on $(O_t, U_t)$ without regard to the treatment effect: $\sum_a Cov(\pi_b(a|o, S), \tau_a(o, S)|o) = 0$. This restricts the behavior policy in aggregate only. Individual actions can still be selected on gains as long as the effects cancel across the action set.*

**Assumption 5** (Persistent Behavior)**.** *We assume that static confounding ensures that $Cov(\pi_b(a|O_t, S_t),\ \pi_b(a|O_{t'}, S_{t'})|o) \geq 0$ for any $t \neq t'$ in the same trajectory.*

**Assumption 6** (Sign fidelity of the history proxy)**.** *We assume that the history $H_t$ is an imperfect proxy for $S_t$, so the covariance we can estimate substitutes $g(o, a, H_t)$ for $m(o, a, S)$ and that the following holds: $sign(Cov(\pi_b(a|o, S), g(o, a, H_t)|o)) = sign(Cov(\pi_b(a|o, S), m(o, a, S)|o))$.*

### 3.3 OUR APPROACH

#### 3.3.1 ZERO-SUM THEORY

As identified in Chernozhukov et al. (2026), the per-$(o, a)$ bias is the covariance of the behavior policy's normalized action probability given the full information set $\frac{\pi_b(a|o,S)}{P(a|o)}$ and the outcome regression $\mathbb{E}[G|o, a, S]$ conditioned on $o$:

$$\text{bias}(o, a) = Q^o(o, a) - Q^c(o, a) = Cov(\frac{\pi_b(a|o, S)}{P(a|o)},\ \mathbb{E}[G|o, a, S]|o) \tag{1}$$

The covariance can be represented in terms of the Pearson correlation $\rho$:

$$\text{bias}(o, a) = \rho \cdot SD(\frac{\pi_b(a|o, S)}{P(a|o)}|o) \cdot SD(\mathbb{E}[G|o, a, S]|o) \tag{2}$$

Chernozhukov et al. (2026) bound the bias of a single estimand. In a sequential decision-making task there are $|\mathcal{A}|$ coupled estimands at each observation, and that coupling constrains the biases to a single scalar.

**Theorem 1** (Zero-sum biases)**.** *Under Assumptions 1, 2 and 4, $\sum_a P(a|o)\,\text{bias}(o, a) = 0$ at every $o$.*

*Proof.*

$$\begin{aligned}
\sum_a P(a|o)\,\text{bias}(o, a) &= \sum_a Cov(\pi_b(a|o, S),\ m(o, a, S)|o) \\
&= \sum_a Cov(\pi_b(a|o, S),\ m(o, a_{ref}, S) + \tau_a(o, S)|o) \\
&= Cov(\sum_a \pi_b(a|o, S),\ m(o, a_{ref}, S)|o) + \sum_a Cov(\pi_b(a|o, S),\ \tau_a(o, S)|o) \\
&= \sum_a Cov(\pi_b(a|o, S),\ \tau_a(o, S)|o) = 0.
\end{aligned}$$

The first line multiplies Eq. 1 by the constant $P(a|o)$. The third moves the sum inside the first covariance, valid because every term shares the second argument. The fourth uses $\sum_a \pi_b(a|o, S) = 1$ for every $S$, so that term is the covariance of a constant. The last is by Assumption 4. □

From above, the weighted biases sum to the aggregate selection-on-gains term regardless of the outcome model, the propensity model, or the degree of effect heterogeneity. Assumption 4 is exactly the condition under which this term vanishes.

Intuitively, confounding by severity mis-ranks the actions at each observation rather than pushing the value of all actions down. For a given $o$, those with worse $\mathrm{sev}$ had non-reference actions taken more often, pushing the observational values of those actions *down*. In contrast, at the same $o$, those with better $\mathrm{sev}$ had more reference actions taken, pushing the observational values of those reference actions *up*. Pessimistic methods give updated Q values of the form $\hat{Q}(o,a) = Q^o(o,a) - \mathrm{pen}(o,a)$ where the penalty satisfies $\mathrm{pen}(o,a) \geq 0$. The difference between the pessimistic-Q and the causal Q value is an error in estimating the causal quantity that we care about, $\mathrm{err}(o,a) = Q^o(o,a) - \mathrm{pen}(o,a) - Q^c(o,a) = \mathrm{bias}(o,a) - \mathrm{pen}(o,a)$.

As we showed above, under our assumptions the propensity-weighted average bias over all actions at any given state is 0. Taking the propensity-weighted average over the given observation $o$ gives $\sum_a P(a|o) \cdot \mathrm{err}(o,a) = \sum_a P(a|o) \cdot \mathrm{bias}(o,a) - \sum_a P(a|o) \cdot \mathrm{pen}(o,a)$. With $\sum_a P(a|o) \cdot \mathrm{bias}(o,a) = 0$, $\mathrm{pen}(o,a) \geq 0$ and $\mathrm{pen}(o,a') > 0$ for at least one $a' \in \mathcal{A}$, the propensity-weighted error at any $o$ must be $< 0$. Moreover, since the biases are zero-sum and not identically zero, at every $o$ there is an action $a^-$ with $\mathrm{bias}(o,a^-) < 0$, which the logged data already under-values. Wherever the penalty is positive at such an action, $\mathrm{err}(o,a^-) = \mathrm{bias}(o,a^-) - \mathrm{pen}(o,a^-) < \mathrm{bias}(o,a^-) < 0$, so the penalized estimate is strictly further from $Q^c$ than the uncorrected one. Sensitivity models with $\Lambda > 1$ lower-bound every cell and therefore have penalties positive at $a^-$ by construction, and CQL's penalty is largest where support is thinnest, which under CBS is where the beneficial action lies.

Additionally, any penalty that is constant across actions at a fixed $o$ has a greedy action that matches the confounded observational policy: $\arg\max_a\{Q^o(o,a) - \mathrm{pen}(o)\} = \arg\max_a\{Q^o(o,a)\}$. This rules out re-ranking only for penalties that do not vary with the action. A penalty that does vary with the action can potentially re-rank actions. However this requires the penalty to vary in the direction that tracks the unobserved confounder, whereas the penalties in existing methods vary with support or model disagreement.

### 3.3.2 Zero-Sum Omitted-variable-bias Corrections (ZSOC)

We use the Chernozhukov et al. (2026) OVB decomposition to construct a correction to the Q-value target during policy training. Regressing the learner on the corrected return targets allows the deployed policy to use the fitted values directly without requiring additional computation during deployment. When constructing the correction, we assume $|\rho(o,a)| = 1$ and estimate the sign separately. This gives a correction motivated by the bias bound rather than an identified estimate of the exact bias magnitude.

We first compute a raw bias correction term $\xi_{raw}(o,a) = \mathrm{dir}(o,a) \cdot \mathrm{out}(o) \cdot \frac{\mathrm{sel}(o,a)}{\hat{P}(a|o)}$. We center the correction to impose the zero-mean constraint under the estimated action probabilities. Centering preserves all pairwise correction contrasts: $\xi(o,a) = \xi_{raw}(o,a) - \sum_{a'} \hat{P}(a'|o) \cdot \xi_{raw}(o,a')$. The learner $\pi$ then regresses on the corrected return, $y_t = G_t - \xi(O_t, A_t)$.

To compute $\xi_{raw}(o,a)$, we must estimate the two quantities defined in the OVB decomposition, the variation in the behavior's action probability due to the unknown confounder $\mathrm{sel}(o,a) = SD(\pi_b(a|o,S)|o)$ and the variation in the outcome due to the unknown confounder $\mathrm{out}(o) = SD(m(o,a,S)|o)$. Additionally, we must estimate a per-$(o,a)$ direction for the correction $\mathrm{dir}(o,a) = sign(\hat{c}(o,a))$, as the product of any two standard deviations is always positive.

First, fit a model $\hat{P}(a|o)$ from the logged data. Next, compute $\mathrm{sel}(o,a) = SD(\pi_b(a|o,S)|o)$. We cannot observe $\pi_b(a|o,S)$, but each logged action is one draw from it. Define the action surprise as $\mathrm{sur}_t(a) = \mathbb{1}[A_t = a] - \hat{P}(a|O_t)$, or the difference between the noisy observed action and the expected action. Because the draw noise is independent across timepoints while $\mathrm{sev}$ is shared, multiplying the surprise against the same trajectory's earlier surprises cancels the noise and leaves the persistent component. Additional details can be found in Appendix A: $\mathrm{sel}(o,a) = \sqrt{\max\{0, \hat{\mathbb{E}}[\mathrm{sur}_t(a) \cdot \mathrm{sur}_{t'}(a)|o]\}} \qquad t' < t.$

By Assumption 5, the lagged product is non-negative and by Cauchy-Schwarz it is bounded by $Var(\pi_b(a|o,S)|o)$, so the estimate is attenuated toward zero and never inflated. The clip handles sampling noise that may erroneously drive the calculated quantity negative.

Then compute $\mathrm{out}(o)$. Under Assumption 3, $Var(m(o,a,S)|o) = Var(m(o,a_{ref},S)|o)$ because $\tau_a(o)$ is a constant given $o$. We therefore use an action-independent estimate of the outcome spread to measure $\mathrm{out}(o)$.

Let $H_t = \{O_{0:t-1}, A_{0:t-1}\}$ and $g(o,a,H_t) = \mathbb{E}[G_t|O_t = o, A_t = a, H_t] - \mathbb{E}[G_t|O_t = o, A_t = a]$. We fit a shared GRU to estimate both conditional means when computing $g(o,a,H_t)$. During training, we zeroed the history with $0.50$ probability to ensure balanced learning for the two conditional means and improve estimator performance. We regress $g(O_t,A_t,H_t)^2$ on $o$ to get $\hat{\sigma}^2(o)$ and set $\mathrm{out}(o) = \sqrt{\max\{\hat{\sigma}^2(o), 0\}}$.

As in Lipsitch et al. (2010); Miao et al. (2018), we assume that the history $H_t$ is a proxy for the state $S_t$, and that it carries enough information to model the effect of $S_t$ on the outcome. This assumption is often violated in practice, as the history may not perfectly represent $S_t$. When this occurs, $g(o,a,H_t) \to 0$, and our estimated $\mathrm{out}(o) \to 0$, reducing the correction and softly degrading ZSOC to an uncorrected Monte Carlo estimator of the outcome.

Finally, compute the sign. The OVB decomposition represents the bias as the product of two standard deviations, resulting in a quantity that is always positive. As such, we must determine the sign of the correction from the data for each $(o,a)$. To do so we read the sign from the covariance between the action surprise and the history gap.

Let $c(o,a) = Cov(\pi_b(a|o,S), g(o,a,H_t)|o)$. Intuitively this tells us if the behavior policy's additional unexpected propensity for $a$ at $o$ align with better or worse outcomes, as predicted by the history-augmented $g(o,a,H_t)$. A negative correlation indicates that the behavior's increased likelihood of taking the action $a$ is associated with *worse* outcomes (and thus the value for this particular $a$ at this $o$ is *undervalued*). In contrast, a positive correlation indicates that the behavior's increased likelihood of taking the action $a$ is associated with *better* outcomes (and thus the value of this particular $a$ at this $o$ is *overvalued*). Thus, the sign of $c(o,a)$ will provide us with the sign of the correction at $(o,a)$.

For each action $a$, we compute $\mathrm{sur}_t(a) \cdot g(O_t,a,H_t)$ for every logged row with non-empty history and regress it on $o$, giving $\hat{c}(o,a) = \hat{\mathbb{E}}[\mathrm{sur}_t(a) \cdot g(O_t,a,H_t)|o]$. Because $\mathrm{sur}_t(a)$ carries the action draw's own noise $\epsilon_t(a)$, the gap must be evaluated counterfactually at every $a$ rather than at the realized action. Substituting $a$ into the head makes $g(O_t,a,H_t)$ measurable with respect to $(O_t,H_t)$ alone, so $\mathbb{E}[\epsilon_t(a) \cdot g|o] = 0$ and the cross term vanishes. Finally, we take the direction of the correction as the sign of $\hat{c}(o,a)$: $\mathrm{dir}(o,a) = sign(\hat{c}(o,a))$.

Our correction is defined for the logged conditional return $\mathbb{E}[G_t|o_t,a_t]$ and thus we apply it to Monte Carlo return targets without bootstrapping. The TD(0) target involves a different, bootstrapped value and therefore requires a separate bias derivation. Though TD(0) approaches can partially deconfound when the next state $O'$ reveals information about the immediate effects of $A$ on $O$, our correction is derived for, and should be applied to, the MC target rather than the bootstrapping target.

All four of the quantities, $\hat{P}(a|o)$, $\mathrm{sel}(o,a)$, $\mathrm{out}(o)$, and $\hat{c}(o,a)$ are computed before training the final value model. At deployment, the policy selects actions greedily using $Q_\theta(o,a)$ and requires no additional computation.

# 4 Experiments

## 4.1 Baseline Methods

DQN (Mnih et al., 2015) and FQI (Ernst et al., 2005) regress the bootstrapped targets $y_t = R_{t+1} + \gamma \max_{a'}\{\hat{Q}(O_{t+1},a')\}$. The observation-based updates do not generally identify causal action values under hidden confounding. However, if an action's effect is revealed in next state $o_{t+1}$, the learner can potentially adjust for the confounding. For discrete actions, CQL adds a regularizer of the form $\alpha \cdot \mathbb{E}[\log \sum_a \exp(Q(O_t,a)) - Q(O_t,A_t)]$ to the value-learning loss. However,

when a beneficial action has little support in the logged data, the regularization underestimates the value of the action. In settings with rare actions and CBI, this can lead to poor policy performance.

Propensity-based deconfounding methods re-weight using the recorded propensity. We include the doubly robust AIPW estimator (Robins et al., 1994), which replaces the target with $\hat{Q}(O_t, a) + \frac{\mathbb{1}[A_t=a]}{\hat{P}(a|O_t)}(G_t - \hat{Q}(O_t, a))$. Its adjustment is a function of $o$ alone, so it cannot remove bias from a confounder the logged data does not contain.

Delphic Offline RL (Delphic ORL) (Pace et al., 2024) estimates uncertainty arising from variation across latent models compatible with the observed data and uses the uncertainty as a pessimistic penalty for policy learning. Confounding-Robust FQI (CR-FQI) (Bruns-Smith & Zhou, 2023) constructs pessimistic Bellman targets under a specified sensitivity model. It provides guarantees on the optimality of the learned policy, but assumes that the confounder is independent across decision points conditional on the context provided to the model. Our confounder is static and affects every step of the trajectory. We apply the fix suggested by Bruns-Smith & Zhou (2023) and provide the method with the previous eight $(o, a)$ pairs to condition on. Both CR-FQI and Delphic ORL make use of user-specified sensitivity parameters that adjust the method based on the assumed strength of confounding. Neither work provides a rule for setting the parameter.

### 4.2 SIMULATORS

We compare ZSOC against the baselines on three environments: a linear Gaussian **Toy** environment, a modified two-way deconfounding environment from Yu et al. (2024), and an ICU simulator with dynamics and outcome hazards fitted to the MIMIC-IV (Johnson et al., 2024; 2023; Pollard et al., 2026) dataset. All three environments follow the SCM of Section 3.1, with the behavior policy taking a uniformly random action with probability $0.10$ to ensure positivity. Each environment also carries an observed binary responder phenotype $B$ drawn once per trajectory that sets the sign of the treatment effect.

An environment with confounding is only challenging if the logged data is misleading about the ranking of actions. As such, we enforced the following conditions in each environment: *The observational ranking of actions is wrong:* For a fixed $o$, the ranking of actions by the observational value $Q^o(o, a)$ must disagree with the ranking by the causal value $Q^c(o, a)$ at a large proportion of $o$ in the logged data. In our environments, no-treatment ranks best at $65 - 97\%$ of the observations where treating is causally better, and the two rankings disagree at $40 - 56\%$ of all observations. *Action effects are not revealed in the next state:* $\mathbb{E}[O_{t+1}|S_t, A_t]$ is independent of $A_t$. Otherwise, if the effects of actions are immediately visible in the next state, TD(0) methods that have access to the next state are able to read part of the treatment effect through $O_{t+1}$ and partially deconfound the bootstrap term. *No constant policy is competitive:* This requirement ensured that a naive learner that simply pushed the values of all actions up was not optimal. Instead, the optimal policy must learn to target, and know when to withhold treatment.

We constructed a Toy environment such that each quantity described in our theory was available in closed form and so that the two axes our claims depend on could be swept independently. We set $S \in \mathbb{R}^6$, where each non-static dimension follows a separate linear-Gaussian autoregressive process. A single multiplier $s$ scales both pathways through which the confounder acts: the loading $\beta_U$ on the read $U_t$ in the behavior policy's logit, and the coefficient $\kappa_{\text{sev}}$ with which $\text{sev}$ drags the transition. Setting $s = 1$ recovers the default operating point.

Yu et al. (2024) provide a process with a static per-trajectory latent, which we take as $\text{sev}$ and which emits a noisy $U$, and a per-time latent $W_t$ shared across trajectories. We make the following changes so it meets our SCM: The original features become the hidden state, and we emit a noisy read of them as $O$. The behavior policy selects on $(O_t, U_t)$ rather than on the state. We remove the original environment's direct action-reward term so that benefit must flow through transitions. We flip the reward's dependence on $\text{sev}$, as in the original environment, a higher $U$ drives both more treatment and better outcomes, which is the opposite of CBS. $W_t$ is retained in the transition where it is shared process noise but removed from the logit and reward so as not to induce confounding on $W_t$.

We fit an ICU environment to $2500$ MIMIC-IV admissions to model start state distributions, transition dynamics, a vasopressor administration policy, and death and discharge hazards. Transition dynamics are fitted only on buckets in which no treatment was given, so the fitted model carries no

treatment effect and the causal effect is the one we inject. The observation is 10 vital signs and the responder phenotype $B$, while the state includes an additional 22 laboratory features. We follow a sparse terminal-only reward model of $-1$ on in-hospital death, $+1$ on discharge, and 0 on truncation. The behavior policy is the fitted vasopressor policy plus a $\beta_U U_t$ term to enable confounding. Appendix B gives the fitting pipeline and the injected constants.

## 5 RESULTS

All fitted methods are evaluated by Monte Carlo rollout of the policy using the environment simulator. For all results, confounding strength is relative to the default configuration on the environment, and does not represent an absolute quantity. We report the results by sweeping over the confounding strength on each environment, rather than reporting a single number, as the default confounding strength in each environment is tuned to ensure that the environment criteria in Section 4.2 are met.

We include behavior cloning (BC) and a neural network fit to estimate the Monte Carlo return from $o, a$ pairs (MC) as controls. In particular, MC serves as the matched control for ZSOC, as it uses the same architecture and training configuration, differing only in that it does not apply our correction.

We measure CR-FQI at two fixed sensitivity values $\Lambda = 2$ and $\Lambda = 20$, which spans mild to severe pessimism. The best performing $\Lambda$ value depends on the confounding, and without oracle knowledge of the environment, is challenging to set in practice. Delphic ORL uses the single configuration ($\lambda = 0.1$).

We find that uncorrected value learning under-treats compared to the oracle policy on all three of our environments. At the default configurations for Toy and Yu, every baseline finishes below the best constant action policy. The separation is clear when considering the treatment rate rather than the value, as ZSOC shows treatment rates of $0.54$ on Toy (vs. $0.53$ Oracle), $0.57$ on Yu (vs. $0.57$ Oracle), and $0.70$ on MIMIC (vs. $0.50$ Oracle). Every baseline settles on a near-constant policy; ours is the only one that conditions the decision on the patient. The pessimistic arms collapse toward never-treating even under weak confounding, while ZSOC moves value between actions and reaches a treatment rate near the oracle's.

At the weakest confounding, our correction is indistinguishable from the uncorrected Monte Carlo learner and from AIPW on all three environments. This is the intended behavior rather than a failure. When there is little confounding to find, both estimated factors shrink and the correction goes to zero, so the method degrades to be an uncorrected Monte Carlo estimator rather than adding noise. The pessimistic methods do not degrade gracefully, and CR-FQI and Delphic ORL are the worst performing methods at low confounding on all three environments.

At intermediate confounding, ZSOC is the best arm everywhere. On Yu it leads at every strength we tested, with the uncorrected one-step arms (DQN, CQL) in the middle and the two pessimistic arms at the bottom. On MIMIC, the spread between arms is narrow throughout, and at intermediate confounding strengths, the ordering behind ZSOC is DQN, then Delphic ORL, then CR-FQI.

At the strongest confounding, we see ZSOC decline in performance while CR-FQI and Delphic ORL improve. On Toy, CR-FQI at $\Lambda = 20$ becomes the best arm. On Yu, ours still leads at $s = 5$, but we observe the margin beginning to narrow. On MIMIC every method collapses to never-treat.

As with other work on confounding in offline settings, our results must be considered cautiously as we show improvements on environments of our own design. In practice, there is no access to the data generating process and so true policy values cannot be determined ahead of time. As such, our results may not transfer to all settings. More work is required to demonstrate the robustness of ZSOC to other types of confounding or to other environments with different dynamics. Finally, we see that as confounding becomes extreme, our policy degrades as the component estimates we compute for the OVB decomposition become more influenced by noise. As such, ZSOC would benefit from alternative approaches to better estimate the OVB component quantities in the presence of extreme confounding.

In summary, we showed that the propensity-weighted per-action biases sum exactly to the aggregate selection-on-gains term, and therefore vanish whenever the behavior policy does not select on gains. It follows that a non-negative penalty has strictly negative weighted-mean error, is strictly worse than no correction at any under-valued action on which it is positive, and cannot re-rank actions at

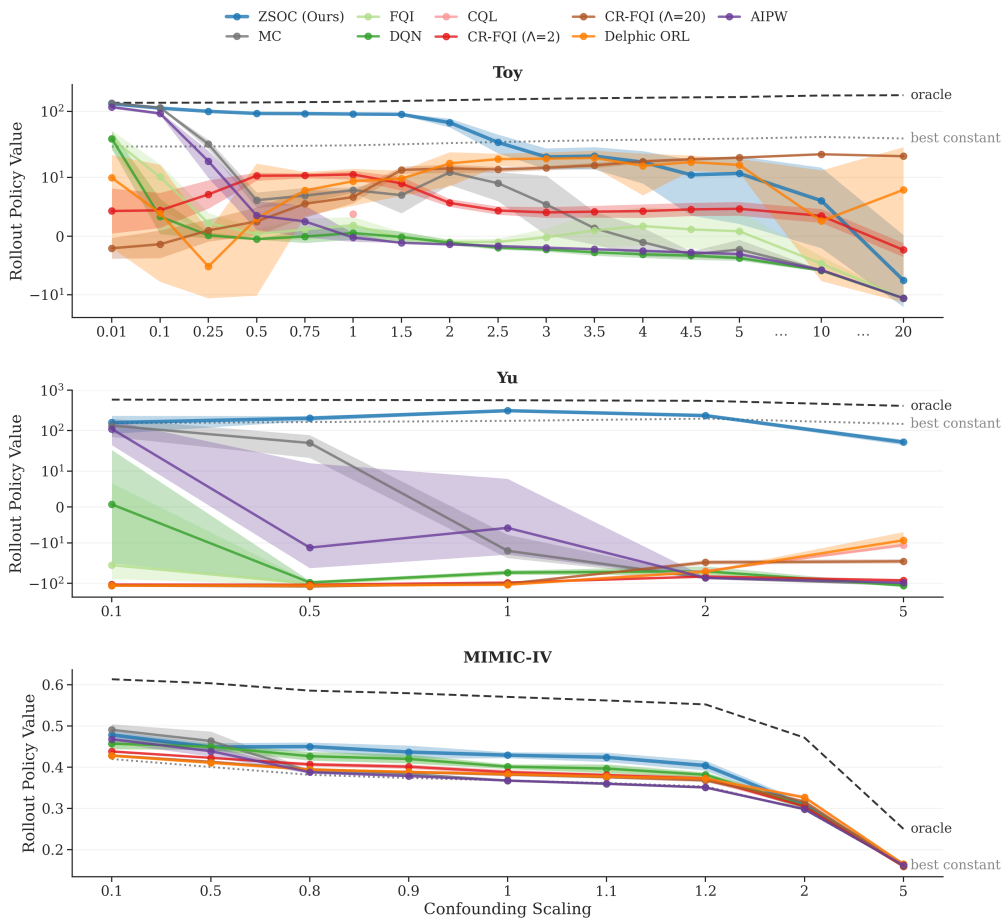



Figure 2: Rollout value against confounding strength $s$ on Toy (top), Yu (middle) and MIMIC-IV (bottom). Confounding scaling, relative to the environment's default strength, is given on the x-axis and policy return is shown on the y-axis. Points indicate mean policy value over five seeds, with shaded bands indicating $\pm 1$ SD. Dashed lines mark the best constant action and the oracle.

all when it does not vary with the action. Additionally, we showed that under a static confounder both factors of the omitted-variable-bias decomposition are estimable from the panel structure of the logged data alone. Finally, we introduced a method that applies a signed zero-sum correction, outperforming all baselines across a range of confounding strengths on three environments with confounding by indication.

## AI USE STATEMENT

In this work, we used generative AI tools to provide critical ingredients for proving mathematical claims, assist in the writing of proofs, and implement methods. We have not used generative AI tools to generate synthetic data sets, help develop theoretical models or conceptual frameworks, propose or refine hypotheses, design or provide feedback on research methodology or experiments, assist with translation, clean and reformat datasets, support qualitative and thematic data analysis, or interpret results. Additionally, we used generative AI tools to create software code, and to revise and edit the research paper to improve readability. We have reviewed all AI-assisted work. Steps in the proofs suggested by AI were checked for correctness by the authors. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

### Reproducibility statement

To facilitate reproducibility, our method is described in detail in Section 3. The configurations for our experiments can be found in Appendix D. The exact structural equations for our environments are provided in Appendix E to ensure the results can be reproduced.

## References


Andrew Bennett and Nathan Kallus. Proximal reinforcement learning: Efficient off-policy evaluation in partially observed markov decision processes. *Operations Research*, 72(3):1071–1086, 2024.

Ioana Bica, Ahmed M Alaa, James Jordon, and Mihaela van der Schaar. Estimating counterfactual treatment outcomes over time through adversarially balanced representations. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=BJg866NFvB.

Donald A Brand, Patricia A Patrick, Jeffrey T Berger, Mediha Ibrahim, Ajsza Matela, Shweta Upadhyay, and Peter Spiegler. Intensity of vasopressor therapy for septic shock and the risk of in-hospital death. *Journal of pain and symptom management*, 53(5):938–943, 2017.

David Bruns-Smith and Angela Zhou. Robust fitted-q-evaluation and iteration under sequentially exogenous unobserved confounders. In *NeurIPS 2023 Workshop on Adaptive Experimental Design and Active Learning in the Real World*, 2023. URL https://openreview.net/forum?id=BctsZxNsfO.

Victor Chernozhukov, Carlos Cinelli, Whitney K Newey, Amit Sharma, and Vasilis Syrgkanis. Long story short: Omitted variable bias in causal machine learning. *The Review of Economics and Statistics*, pp. 1–45, 2026.

Zhihong Deng, Jing Jiang, Guodong Long, and Chengqi Zhang. Causal reinforcement learning: A survey. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856. URL https://openreview.net/forum?id=qqnttX9LPo. Survey Certification.

Damien Ernst, Pierre Geurts, and Louis Wehenkel. Tree-based batch mode reinforcement learning. *Journal of Machine Learning Research*, 6(18):503–556, 2005. URL http://jmlr.org/papers/v6/ernst05a.html.

Omer Gottesman, Fredrik Johansson, Matthieu Komorowski, Aldo Faisal, David Sontag, Finale Doshi-Velez, and Leo Anthony Celi. Guidelines for reinforcement learning in healthcare. *Nature medicine*, 25(1):16–18, 2019.

Tobias Hatt and Stefan Feuerriegel. Sequential deconfounding for causal inference with unobserved confounders. In Francesco Locatello and Vanessa Didelez (eds.), *Proceedings of the Third Conference on Causal Learning and Reasoning*, volume 236 of *Proceedings of Machine Learning Research*, pp. 934–956. PMLR, 01–03 Apr 2024. URL https://proceedings.mlr.press/v236/hatt24a.html.

Alistair Johnson, Lucas Bulgarelli, Tom Pollard, Brian Gow, Benjamin Moody, Steven Horng, Leo Anthony Celi, and Roger Mark. MIMIC-IV. *PhysioNet*, October 2024. doi: 10.13026/kpb9-mt58. Version 3.1.

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

Aviral Kumar, Aurick Zhou, George Tucker, and Sergey Levine. Conservative q-learning for offline reinforcement learning. *Advances in neural information processing systems*, 33:1179–1191, 2020.

Demetrios N. Kyriacou and Roger J. Lewis. Confounding by indication in clinical research. *JAMA*, 316(17):1818–1819, 11 2016. ISSN 0098-7484. doi: 10.1001/jama.2016.16435.

Sergey Levine, Aviral Kumar, George Tucker, and Justin Fu. Offline reinforcement learning: Tutorial, review, and perspectives on open problems. *arXiv preprint arXiv:2005.01643*, 2020.

Luofeng Liao, Zuyue Fu, Zhuoran Yang, Yixin Wang, Dingli Ma, Mladen Kolar, and Zhaoran Wang. Instrumental variable value iteration for causal offline reinforcement learning. *Journal of Machine Learning Research*, 25(303):1–56, 2024. URL http://jmlr.org/papers/v25/22-0965.html.

Marc Lipsitch, Eric Tchetgen Tchetgen, and Ted Cohen. Negative controls: a tool for detecting confounding and bias in observational studies. *Epidemiology*, 21(3):383–388, May 2010.

Christos Louizos, Uri Shalit, Joris M Mooij, David Sontag, Richard Zemel, and Max Welling. Causal effect inference with deep latent-variable models. *Advances in neural information processing systems*, 30, 2017.

Chaochao Lu, Bernhard Schölkopf, and José Miguel Hernández-Lobato. Deconfounding reinforcement learning in observational settings. *arXiv preprint arXiv:1812.10576*, 2018.

Miao Lu, Yifei Min, Zhaoran Wang, and Zhuoran Yang. Pessimism in the face of confounders: Provably efficient offline reinforcement learning in partially observable markov decision processes. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=PbkBDQ5_UbV.

Valentyn Melnychuk, Dennis Frauen, and Stefan Feuerriegel. Causal transformer for estimating counterfactual outcomes. In *International conference on machine learning*, pp. 15293–15329. PMLR, 2022.

Wang Miao, Zhi Geng, and Eric Tchetgen Tchetgen. Identifying causal effects with proxy variables of an unmeasured confounder. *Biometrika*, 105(4):987–993, December 2018.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, et al. Human-level control through deep reinforcement learning. *nature*, 518(7540):529–533, 2015.

Hongseok Namkoong, Ramtin Keramati, Steve Yadlowsky, and Emma Brunskill. Off-policy policy evaluation for sequential decisions under unobserved confounding. *Advances in Neural Information Processing Systems*, 33:18819–18831, 2020.

Alizée Pace, Hugo Yèche, David Ha, Gunnar Ratsch, and Guy Tennenholtz. Delphic offline reinforcement learning under nonidentifiable hidden confounding. In *International Conference on Learning Representations*, volume 2024, pp. 2768–2796, 2024.

Judea Pearl. Causal inference in the health sciences: a conceptual introduction. *Health services and outcomes research methodology*, 2(3):189–220, 2001.

Judea Pearl. Causal inference. *Causality: objectives and assessment*, pp. 39–58, 2010.

Tom Pollard, Benjamin E. Moody, Li-wei H. Lehman, Brian J. Gow, Chrystinne Fernandes, Chen Xie, Alistair Johnson, Roger G. Mark, and Thomas Heldt. PhysioNet as a global platform for biomedical research. *Nature Health*, 1(8):792–795, August 2026. ISSN 3005-0693. doi: 10.1038/s44360-026-00096-z.

James M Robins, Andrea Rotnitzky, and Lue Ping Zhao. Estimation of regression coefficients when some regressors are not always observed. *Journal of the American statistical Association*, 89 (427):846–866, 1994.

Paul R Rosenbaum. Overt bias in observational studies. In *Observational studies*, pp. 71–104. Springer, 2002.

Chengchun Shi, Masatoshi Uehara, Jiawei Huang, and Nan Jiang. A minimax learning approach to off-policy evaluation in confounded partially observable markov decision processes. In *International Conference on Machine Learning*, pp. 20057–20094. PMLR, 2022.

Zhiqiang Tan. A distributional approach for causal inference using propensity scores. *Journal of the American Statistical Association*, 101(476):1619–1637, 2006.

Eric J Tchetgen Tchetgen, Andrew Ying, Yifan Cui, Xu Shi, and Wang Miao. An introduction to proximal causal learning. *medRxiv*, 2020. doi: 10.1101/2020.09.21.20198762.

Eric J Tchetgen Tchetgen, Andrew Ying, Yifan Cui, Xu Shi, and Wang Miao. An introduction to proximal causal inference. *Statistical Science*, 39(3):375–390, 2024.

Anpeng Wu, Kun Kuang, Ruoxuan Xiong, and Fei Wu. Instrumental variables in causal inference and machine learning: A survey. *ACM Computing Surveys*, 57(11):1–36, 2025.

Chao Yu, Jiming Liu, Shamim Nemati, and Guosheng Yin. Reinforcement learning in healthcare: A survey. *ACM Computing Surveys (CSUR)*, 55(1):1–36, 2021.

Shuguang Yu, Shuxing Fang, Ruixin Peng, Zhengling Qi, Fan Zhou, and Chengchun Shi. Two-way deconfounder for off-policy evaluation in causal reinforcement learning. *Advances in Neural Information Processing Systems*, 37:78169–78200, 2024.

Yan Zeng, Ruichu Cai, Fuchun Sun, Libo Huang, and Zhifeng Hao. A survey on causal reinforcement learning. *IEEE Transactions on Neural Networks and Learning Systems*, 36(4):5942–5962, 2025. doi: 10.1109/TNNLS.2024.3403001.

Qingyu Zhao, Ehsan Adeli, and Kilian M Pohl. Training confounder-free deep learning models for medical applications. *Nature communications*, 11(1):6010, 2020.

# A Additional Derivations

## A.1 Derivation of $\mathrm{sel}(o, a)$

The realized action is a noisy draw from the behavior policy, $\mathbb{1}[A_t = a] = \pi_b(a|O_t, S_t) + \epsilon_t(a)$. We define the action surprise $\mathrm{sur}_t(a) = \mathbb{1}[A_t = a] - \hat{P}(a|O_t) = \pi_b(a|O_t, S_t) - \hat{P}(a|O_t) + \epsilon_t(a)$, with $\mathbb{E}[\epsilon_t(a)|O_t, S_t] = 0$.

Thus $\mathbb{E}[\mathrm{sur}_t(a) \cdot \mathrm{sur}_{t'}(a)|o] = \mathbb{E}[(\pi_b(a|O_t, S_t) - \hat{P}(a|O_t)) \cdot (\pi_b(a|O_{t'}, S_{t'}) - \hat{P}(a|O_{t'}))|o]$.

By the definition of covariance this equals $Cov(\pi_b(a|O_t, S_t), \pi_b(a|O_{t'}, S_{t'})|o)$. Within a trajectory, any two action surprises share the same $\mathrm{sev}$, so their product has non-zero mean. However, the two timepoints for which the surprises are computed carry different states, and so the covariance across time under-estimates the within-chart variance.

We estimate $\mathrm{sel}(o, a) = \sqrt{\mathbb{E}[\mathrm{sur}_t(a) \cdot \mathrm{sur}_{t'}(a)|o]}$ over all $t' \in \{0, \ldots, t-1\}$. By Assumption 5, the lagged product is non-negative, and by Cauchy-Schwarz, $Cov(\pi_b(a|O_t, S_t), \pi_b(a|O_{t'}, S_{t'})|o) \leq Var(\pi_b(a|o, S)|o)$, so our estimate is never inflated.

# B Environment Details

All three environments implement the same causal graph with different structural equations.

Let $S_t = (X_t, \mathrm{sev}, \mathrm{eff}_t)$, where $X_t$ is the physiology block that is used to emit $O$, $\mathrm{sev}$ is the static confounder dimension of $S$, and $\mathrm{eff}_t$ is a set of treatment-affected dimensions. All environments satisfy the following:

1. $S_t \to S_{t+1} \leftarrow A_t$, with $\dim(S) = k \geq 5$.
2. $O_t = X_t[\mathcal{C}] + \sigma_O \,\mathrm{sd}(X)\, \varepsilon_t$.
3. $U_t = \mathrm{sev} + \sigma_U \varepsilon_t$.
4. $O_t$ and $U_t$ read disjoint coordinates of $S_t$.
5. The learner sees $O_t$ alone.
6. $A_t \sim \pi_b(\cdot|O_t, U_t)$, with a $10\%$ uniformly random action.
7. $R_{t+1} = r(S_{t+1})$.
8. $\mathbb{E}[O_{t+1}|S_t, A_t]$ is invariant to $A_t$.
9. $A_t$ never reaches $\mathrm{sev}$, which no action can move. This is what makes the confounding static.
10. Both emissions are noisy: $\sigma_O > 0$ and $\sigma_U > 0$.

On Toy and Yu, the dose enters a compartment and does not reach the physiology for one step, so $\mathbb{E}[O_{t+1}|S_t, A_t]$ is identical across actions. On MIMIC-IV the dose acts within the same bucket, but along a direction with exactly zero support in $O$, so it is not revealed in the next observation. The treatment effect is a function of $O$, not the hidden state, so it is constant given $o$ and the oracle policy is reachable by a learner that sees $o$ alone.

**Fitting MIMIC-IV** To fit the MIMIC-IV environment, we draw $2500$ admissions from MIMIC-IV and bin each into $4$-hour buckets with forward-fill imputation to replace missing values. We select $10$ vital signs and $22$ laboratory measurements as features.

We fit the following:

1. the $32$ features and their scales.
2. The episode-start feature distribution.
3. The natural drift for consecutive buckets in which no treatment action was given. Fit with a two-layer $64$-unit diagonal-Gaussian MLP. Restricting to untreated transitions is what keeps the fitted dynamics free of any treatment effect, enabling us to define our own causal effects.

4. The death axis $w_{\text{death}}$, which is a logistic regression of the per-bucket death on the state (held-out AUC $0.915$) and is used to set the structural direction.
5. The real vasopressor policy. Fit by a logistic regression of the per-bucket vasopressor indicator on the vitals (held-out AUC $0.856$, rate $0.066$);
6. The death and discharge models. Both fit with a two-layer $64$-unit MLP to predict death (AUC $0.931$) and discharge (AUC $0.784$).

**What is injected, and why.** A benchmark needs a causal effect that is *known*, and an effect read off confounded observational data would not be one. Table 1 lists every injected constant.

Table 1: The injected parameters for the MIMIC-IV environment. All other parameters are set by fitting to the data.

| symbol | value | role |
|---|---|---|
| $\sigma_{\text{sev}}$ | $1.00$ | $\text{sev} \sim \mathcal{N}(0, 1)$, static for the entire trajectory |
| $\sigma_U$ | $0.50$ | the clinician's read noise on $\text{sev}$ |
| $\sigma_O$ | $0.30$ | observation noise |
| $p_B$ | $0.50$ | prevalence of the binary responder phenotype $B$ |
| $\kappa_{\text{treat}}$ | $0.20$ | scalar effect of one dose, in pre-$\tanh$ units |
| $\kappa_{\text{sev}}$ | $0.035$ | scalar effect of $\text{sev}$ |
| $c_B$ | $2.50$ | how much more $\text{sev}$ progresses faster in responders |
| $h$ (harm ratio) | $0.45$ | the non-responder's harm from treatment as a share of the responder's |
| $\beta_U$ | $4.00$ | sets the behavior dependence on $U$ |

Data generated by the simulator reproduced the observed outcome distribution from the dataset, with a rate of death of $0.097$ and discharge of $0.478$, against $0.105$ and $0.507$ in the logged data.

## C Measured Environment Properties

Table 2: Analysis of the observational and causal action value disagreement, measured via Monte Carlo rollout with 3000 samples. "Obs. $\arg\max$ is no-treat" is restricted to observations where treating is causally better.

| Environment | Obs. $\arg\max$ is no-treat | Obs. vs. causal disagree |
|---|---|---|
| Toy | $87.1\%$ | $55.7\%$ |
| Yu | $65.4\%$ | $40.0\%$ |
| MIMIC-IV | $96.6\%$ | $47.0\%$ |

Table 2 shows a comparison of the observational and causal action value disagreement. Yu is the environment where the observational ranking is least uniformly wrong, at $65.4\%$. This is because treatment is so strongly beneficial that the log is misleading about the ranking on a smaller share of observations.

Table 3: Informativeness of the emissions. $R^2(\text{sev}|O)$ is how much of the confounder the chart already reveals; $R^2(\text{sev}|O, U)$ is how much the clinician gains by observing $U$.

| Environment | $R^2(\text{sev}\|O)$ | $R^2(\text{sev}\|O, U)$ |
|---|---|---|
| Toy | $0.592$ | $0.882$ |
| Yu | $0.688$ | $0.900$ |
| MIMIC-IV | $0.280$ | $0.790$ |

Table 3 shows how much of the confounder is revealed by $O$. On Yu and Toy the confounder is largely recoverable from the chart alone ($R^2 = 0.688$ and $0.592$). As such, we cannot claim the confounder is unrecoverable on those two environments.

MSMs bound the odds ratio between the confounded and the observed propensity with a single $\Lambda$ parameter. Table 4 shows the distribution of that ratio in our environments, computed using the true $\pi_b(\text{treat}|o, S)$ by rolling out on the environment.

Table 4: Distribution of the true odds ratio between $\pi_b(\text{treat}|o, S)$ and $P(\text{treat}|o)$, which is the quantity a marginal sensitivity model must cover with one $\Lambda$.

| Environment | Median $\Lambda$ | 90th Pctl. | 99th Pctl. | 99.9th Pctl. | Max |
|---|---|---|---|---|---|
| Toy | 1.95 | 4.98 | 19.6 | 56 | 173 |
| Yu | 1.75 | 4.32 | 31.3 | 98 | 315 |
| MIMIC-IV | 3.35 | 9.79 | 28.5 | 55 | 129 |

# D Method and Baseline Configurations

Every arm runs on the same dataset, network width and depth, learning rate and schedule, batch size, and step budget.

Table 5: Training configuration shared by every method.

| Setting | Value |
|---|---|
| Hidden dim | 256 |
| Layers | 3 |
| Learning rate | $10^{-4}$ |
| Batch size | 4096 |
| Epochs | 250 |
| $\gamma$ | 0.95 |
| Gradient clip | 10.0 |
| FQI iterations | 30 |
| DQN steps before target update | 200 |
| Training patients | 6000 |
| Validation patients | 1500 |

## D.1 Our Configuration

The fitted $\hat{P}(a|o)$ and $\hat{c}(o, a)$ models are gradient-boosted tree ensembles, the history model is a single two-layer GRU with hidden width $64$ trained for $50$ epochs at learning rate $10^{-3}$ with batches of size $1024$. The history representation is zeroed with probability $0.5$ during training.

## D.2 Baseline Configurations

CQL uses $\alpha = 1.0$ on every environment. CR-FQI cross-fits its propensity by patient over $4$ folds and conditions both nuisances on the previous eight $(o, a)$ pairs. The deployed $Q$ remains memoryless to enable fair comparison with all other methods. Delphic ORL uses $5$ world models consisting of a per-step MLP encoder with dims $(128, 64, 32)$, a learning rate of $10^{-3}$, latent dims sampled from $\{1, 2, 4, 8, 16\}$, prior variance sampled from $\{1, 0.1, 0.01\}$, 16 Monte Carlo $z$-draws, 5 patient-level bootstraps per world, a 200-quantile $Q$ head, a softmax policy at temperature $1.0$, and a $10^{-3}$ floor on the per-world behavior probability. For Delphic ORL, we use an MLP model and applied a floor on the per-world behavior probability to prevent errors. We trained the model to convergence instead of using early stopping as described in the paper. Delphic ORL provides no rule for setting its sensitivity parameter, and the value that maximizes true policy value is not available to a practitioner.

# E Full Environment Specifications

$S_t = (X_t, \text{sev}, \text{eff}_t)$. $\varepsilon$ is an independent standard normal, fresh at each use. $d$, $g$, $\rho$ are unit vectors (severity, eligibility, relief) with $g \perp d$. $z(\cdot)$ is the standardized observation. All three environments replace $A_t$ with a uniform draw with probability $0.10$.

## E.1 Toy

1. $\dim(X) = 4$, $k = 6$, $\dim(O) = 4$, $\dim(U) = 1$, $|\mathcal{A}| = 3$, $\text{dose}(a) \in \{0, 1, 2\}$, $T = 30$.

2. $\phi = 0.85$, $\sigma_x = 1.00$, $\kappa_{\text{sev}} = 0.75$, $\rho_{\text{eff}} = 0.75$, $\|\rho\| = 1.00$.
3. $\sigma_{\text{sev}} = 1.00$, $\sigma_U = 0.40$, $\sigma_O = 0.35$.
4. $\beta_U = 2.20$, $\beta_O = -0.60$, $b_0 = -0.55$.
5. $w_g = 2.50$, $c_g = 0$, $\text{tox} = 0.45$.
6. $c_x = 1.00$, $c_{\text{sev}} = 1.50$.
7. $d \propto (1, 1, 1, 1)$, $g \propto (1, -1, 1, -1)$, $\rho \propto (0.9, 1.1, 0.8, 1.2)$.
8. $\text{sev} \sim \mathcal{N}(0, \sigma^2_{\text{sev}})$, drawn once; $\text{eff}_0 = 0$; $X_0^i = \dfrac{-\kappa_{\text{sev}} d_i \text{sev}}{1-\phi} + \dfrac{\sigma_x}{\sqrt{1-\phi^2}}\varepsilon$.
9. $O_t = X_t + \sigma_O \,\text{sd}(X)\,\varepsilon$, $U_t = \text{sev} + \sigma_U \varepsilon$.
10. $\text{proxy}(o) = d \cdot z(o)$, $\text{elig}(o) = g \cdot z(o)$, $e(o) = \text{sigmoid}(w_g(\text{elig}(o) - c_g)) - \text{tox}$.
11. $A_t \sim \text{Binomial}(|\mathcal{A}| - 1,\ \text{sigmoid}(\beta_O \text{proxy}(O_t) + \beta_U U_t + b_0))$.
12. $X^i_{t+1} = \phi X^i_t - \kappa_{\text{sev}} d_i \text{sev} + \rho_i \text{eff}_t + \sigma_x \varepsilon$.
13. $\text{eff}_{t+1} = \rho_{\text{eff}} \text{eff}_t + e(O_t)\,\text{dose}(A_t)$; sev unchanged.
14. $R_{t+1} = c_x(d \cdot X_{t+1}) - c_{\text{sev}} \text{sev}$.

### E.2 YU

1. $\dim(X) = 4$, $k = 6$, $\dim(O) = 4$, $\dim(U) = 1$, $|\mathcal{A}| = 2$, $\text{dose}(a) \in \{0, 1\}$, $T = 50$.
2. Retained from Yu et al. (2024): $\Gamma = 2.0$, $\sigma = 1.0$, $\phi = 0.8$, $\delta = -0.5$, $w_x = 0.25$, $b_{\text{orig}} = -4.0$.
3. Added or changed: $\kappa_{\text{sev}} = 0.80$, $\|\rho\| = 4.00$, $\rho_{\text{eff}} = 0.75$, $\sigma_{\text{sev}} = 1.00$, $\sigma_U = \sigma_O = 0.35$, $w_g = 2.50$, $c_g = 0$, $\text{tox} = 0.45$.
4. $d \propto (1, 1, 1, 1)$, $g \propto (1, -1, 1, -1)$, $\rho \propto (1.1, 0.9, 1.2, 0.8)$.
5. $W \sim \mathcal{N}(0, I_{T+1})$, drawn once at construction and shared by all trajectories; $z_1 = \text{sev} W_t$, $z_2 = \text{sev} - W_t$, $z_3 = \text{sev} + W_t$, $z_4 = -\text{sev} - W_t$, $z_5 = -\text{sev} + W_t$.
6. $\text{sev} \sim \mathcal{N}(0, \sigma^2_{\text{sev}})$, drawn once; $\text{eff}_0 = 0$; $X_0^j = \varepsilon - \dfrac{\kappa_{\text{sev}} d_j \text{sev}}{1-\phi}$.
7. $O_t = X_t + \sigma_O \,\text{sd}(X)\,\varepsilon$, $U_t = \text{sev} + \sigma_U \varepsilon$.
8. $\text{elig}(o) = g \cdot z(o)$, $e(o) = \text{sigmoid}(w_g(\text{elig}(o) - c_g)) - \text{tox}$.
9. $A_t \sim \text{Bernoulli}(\text{sigmoid}(w_x \sum_j O_{t,j} + b_{\text{orig}} + 3\Gamma U_t))$.
10. $X^j_{t+1} = \phi X^j_t + 0.1\Gamma[z_2, z_3, z_4, z_5]_j + \delta - \kappa_{\text{sev}} d_j \text{sev} + \rho_j \text{eff}_t + \sigma\varepsilon$.
11. $\text{eff}_{t+1} = \rho_{\text{eff}} \text{eff}_t + e(O_t)\,\text{dose}(A_t)$; sev unchanged.
12. $R_{t+1} = w_x \sum_j X_{t,j} - 3\Gamma \text{sev} + \sigma\varepsilon$. Following the original process, this reward is evaluated at $X_t$ rather than $X_{t+1}$; it remains a function of the patient's state and noise alone.

### E.3 MIMIC-IV

1. $\dim(X) = 33$ (32 fitted features, plus $B$), $k = 34$, $\dim(O) = 11$, $\dim(U) = 1$, $|\mathcal{A}| = 2$, $\text{dose}(a) \in \{0, 1\}$, $T = 48$ buckets of 4 hours. No compartment.
2. Charted: 10 vitals and $B$. Censored: 22 laboratory features.
3. Features standardized then squashed by $\tanh(z/3)$; pushes applied in $\text{atanh}$ coordinates, clipped to $\pm(1 - 10^{-6})$.
4. $\sigma_{\text{sev}} = 1.00$, $\sigma_U = 0.50$, $\sigma_O = 0.30$, $p_B = 0.50$, $\kappa_{\text{treat}} = 0.20$, $\kappa_{\text{sev}} = 0.035$, $c_B = 2.50$, $h = 0.45$, $\beta_U = 4.00$.
5. Fitted: the 32 features and scales; $\mathbf{z}_{\text{pool}}$ (real first buckets); drift (diagonal-Gaussian MLP over the one-bucket increment, trained on transitions with no action); $w_{\text{death}}$ (logistic, per-bucket death on the state); $(b_0^\pi, w^\pi)$ (logistic, per-bucket vasopressor on the vitals); $\text{haz}_{\text{died}}$, $\text{haz}_{\text{home}}$ (MLPs, per-bucket hazards).

6. $\hat{d} = \mathrm{unit}(w_{\mathrm{death}}|_{\mathrm{labs}})$, zero on all charted coordinates.
7. $\mathrm{sev} \sim \mathcal{N}(0, \sigma^2_{\mathrm{sev}})$, drawn once; $B = +0.5$ w.p. $p_B$, else $-0.5$, fixed for the trajectory; $X_0 = [\mathbf{z}, B]$, $\mathbf{z} \sim \mathbf{z}_{\mathrm{pool}}$.
8. $O_t = [X_t[\mathrm{vitals}],\ B] + \sigma_O\, \mathrm{sd}_{\mathrm{real}}\, \varepsilon, \qquad U_t = \mathrm{sev} + \sigma_U \varepsilon.$
9. $e(O_t) = +1$ if $B > 0$, else $-h$.
10. $A_t \sim \mathrm{Bernoulli}(\mathrm{sigmoid}(b_0^\pi + w^\pi \cdot O_t[\mathrm{vitals}] + \beta_U U_t))$.
11. $X_{t+1} = \tanh(\mathrm{atanh}(X_t + \mathrm{drift}(X_t)) + [\kappa_{\mathrm{sev}}\mathrm{sev}(1 + c_B \mathbb{1}[B > 0]) - \kappa_{\mathrm{treat}} e(O_t)\mathrm{dose}(A_t)]\hat{d})$; $B$, sev unchanged.
12. $R_{t+1} = -1$ w.p. $\mathrm{haz}_{\mathrm{died}}(X_{t+1})$; else $+1$ w.p. $\mathrm{haz}_{\mathrm{home}}(X_{t+1})$; else $0$. Both non-zero outcomes absorbing; truncation at $t = T$.