CauTwin: Counterfactual Contrast Supervision for Interventional Prediction in Clinical Time Series

Anonymous authors
Paper under double-blind review

Abstract

Estimating how a patient would respond to a treatment that was not given is central to clinical decision support, and electronic health records make it possible to learn such responses from observational data. Existing methods for treatment effects over time adjust for time-varying confounding through balanced representations, inverse propensity weighting, G-computation or meta-learners, and have made effect estimates far more reliable. In intensive care, however, two problems remain. First, treatments are given to patients who are already deteriorating, so models confuse this association with the effect of the treatment and may conclude, for example, that a vasopressor lowers blood pressure. Second, each patient is observed under one treatment only, so the data never show how the effect differs between patients, and models tend to give nearly the same answer to everyone. We propose CauTwin, which learns the effect of a treatment on each patient directly. For every patient, it finds similar patients who received the other treatment and uses this comparison to guide the model's own interventional prediction. Because the comparison remains biased on average by severity that the records do not capture, CauTwin uses only how it differs between patients, which recovers how the effect varies across patients, and sets the average effect with adversarial balancing. We show that the two components compete over how the treatment may depend on severity, and that their weights control this trade-off. On three ICU cohorts, CauTwin is the only method whose vasopressor effects both point in the pharmacologically correct direction and vary with the patient, and it attains the lowest counterfactual error on a semi-synthetic tumour growth benchmark. We further show that predictive accuracy is a misleading criterion for selecting such models.

1 Introduction

Predicting deterioration or forecasting vital signs from electronic health records is now routine for supervised sequence models. The question a clinician faces is different: how would *this* patient respond to a treatment that was not given? Answering it from observational data is central to personalized medicine (Komorowski et al., 2018; Hernán & Robins, 2020), and the quantity it asks for is the *interventional contrast*: the difference in outcome under two treatments for one fixed patient history.

Existing methods. A large body of work estimates treatment effects over time from electronic health records. Balancing methods learn a representation of the patient history from which the treatment decision cannot be predicted (Bica et al., 2020b; Melnychuk et al., 2022). Other methods reweight patients by the inverse probability of the treatment they received (Lim et al., 2018), or simulate the future step by step under a chosen treatment through G-computation (Li et al., 2021; Hess et al., 2026a). More recent meta-learners bring doubly robust and orthogonal estimation to this setting (Frauen et al., 2025; Hess et al., 2026b). The latter methods seek stability under weak overlap within the observed support; no method identifies effects where treatment support is absent. Together, these methods have made treatment effects far more reliable to estimate under confounding. Several model-based

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.

methods obtain treatment effects by contrasting predicted trajectories, without an objective on the patient-level contrast itself. Meta-learners do target the effect directly, but fit it with a separate model that need not agree with the model producing the patient's trajectory.

Challenges. The difficulty lies in the data. Each patient is observed under one treatment only, so the data never show how the same patient would respond to both. A model can learn the effect only by comparing different patients, and in the ICU treated and untreated patients differ in systematic ways. Vasopressors are started because blood pressure is already low, and part of the severity behind this decision is seen at the bedside but never recorded. Two problems follow. First, nothing in the data separates this difference between patients from the effect of the treatment, so the average effect, which we call the *level*, can point in the wrong direction: a model may conclude that vasopressors lower blood pressure. Second, how the effect varies across patients, which we call the *shape*, is not shown by the data at all, so a model tends to give nearly the same answer to every patient. What is missing is a direct signal about how each patient would respond to the treatment that was not given.

Our approach. We propose **CauTwin**, which learns the effect of a treatment on each patient directly, rather than as a by-product of predicting observed outcomes (Figure 1). Since no patient is observed under both treatments, we construct the missing comparison. For each patient, we find similar patients who received the other treatment and compare their outcomes. This comparison is not reliable in its average, because similar patients still differ in severity that the records do not capture. It is reliable in how it differs between patients, which is exactly the information the data otherwise lack. Our counterfactual contrast supervision (§4.2) therefore teaches the model only these differences between patients, and we prove that this recovers the shape of the effect without touching its level. Adversarial balancing (§4.3), taken from Bica et al. (2020b), sets the level by preventing the treatment from standing in for severity. The two parts pull in opposite directions: balancing keeps severity out of the treatment pathway, while the size of the effect should depend on severity. This tension cannot be removed by separating the two parts inside the model, and their weights set a trade-off between the direction of the effect and its dependence on the patient.

Contributions. Our contributions are four-fold:

1. **New insight.** We show that learning from observed outcomes alone leaves the effect of a treatment on each patient undetermined, and that this explains why such models give effects of the wrong direction that barely depend on the patient (§3).
2. **New method.** We propose *counterfactual contrast supervision*, which guides the effect produced by the same model that predicts the patient's trajectory, using comparisons with similar patients who received the other treatment. We show what this guidance recovers and how it interacts with balancing (§4).
3. **Empirical performance.** On three ICU cohorts, CauTwin is the only method whose effects both point in the correct direction and depend on the patient. On the tumour benchmark it attains the lowest counterfactual error at every confounding level (§5).
4. **A warning for practice.** Predictive accuracy misleads model selection: likelihood and interventional correctness favour opposite ends of the same sweep (§5.4).

2 Related work

Treatment effects over time. Estimating treatment effects from longitudinal observational data builds on the G-methods of Robins (1986); Robins et al. (2000), and neural methods follow the same families (Hernán & Robins, 2020). Balancing methods (CRN (Bica et al., 2020b), CT (Melnychuk et al., 2022), ACTIN (Wang et al., 2024)) learn a history representation from which the treatment cannot be predicted. Representation balancing was developed for counterfactual prediction under covariate shift (Johansson et al., 2016; Shalit et al., 2017). It may introduce bias if the representation drops confounders (Melnychuk et al., 2024); temporal balancing has also been examined empirically (Huang et al., 2024). RMSN (Lim et al., 2018) reweights patients by the inverse probability of their treatment, which becomes unstable over long horizons as these probabilities multiply. Pareto smoothing

addresses related instability in inverse-propensity weighting for static heterogeneous effects (Chikahara & Ushiyama, 2024). G-Net (Li et al., 2021) and IGC-Net (Hess et al., 2026a) perform G-computation, the latter replacing Monte Carlo simulation of future covariates with iterated regression, and TE-CDE (Seedat et al., 2022) extends the setting to continuous time. Deep longitudinal TMLE uses a transformer and targeted correction for dynamic policies (Shirakawa et al., 2024). These methods adjust for time-varying confounding under identification assumptions. We address a complementary question, namely how to guide the model on how the effect differs between patients, and we reuse balancing from this line as one component of our method. Causal CPC uses contrastive predictive coding for longitudinal representation learning (El Bouchattaoui et al., 2024), whereas our contrast loss supervises the predicted patient-level treatment difference.

Meta-learners. Meta-learners target the effect directly. The DR-learner (Kennedy, 2023) and the R-learner (Nie & Wager, 2021) construct pseudo-outcomes whose expectation is the effect and fit a second model to them. Weight choice affects the stability of such regressions (Fisher, 2024). Time-varying extensions include model-agnostic pseudo-outcome learners (Frauen et al., 2025) and weighting toward supported intervention sequences (WO-learner; Hess et al., 2026b) and orthogonal estimation for Markov decision processes (DRQ-learner; Javurek et al., 2026). These methods come with strong guarantees when the recorded data capture all confounding. Our approach differs in two ways. First, we do not fit a separate effect model: the comparison guides the effect of the same model that predicts the trajectory, so the predicted trajectories and the predicted effects agree. Second, our comparison comes from simple propensity matching (Rosenbaum & Rubin, 1983; Austin, 2011), and we use only how it differs between patients, because in the ICU its average remains biased by unrecorded severity. This trades efficiency for robustness to a bias shared across patients.

Learned models for clinical decisions. Offline reinforcement learning for sepsis (Komorowski et al., 2018) relies on a learned model of how patients evolve under treatment, whose counterfactual predictions cannot be checked against data, a concern raised by Gottesman et al. (2019) and Jeter et al. (2019). Such models are usually selected by how well they predict the observed future. We examine this practice on multi-centre ICU data (Johnson et al., 2023; Pollard et al., 2018) and find that it favours models whose interventional answers point in the wrong direction (§5.4). This connects to work on counterfactual model selection (Curth & van der Schaar, 2023; Gao, 2025). An extended review is in Appendix G.

3 Problem Formulation

Setup. At each hour t, a patient has recorded physiological values X_t, a mask M_t that marks which values were measured, and a treatment indicator A_t. We write H_t for the patient's history up to hour t. A model predicts the next hour from the history and the current treatment, and is rolled forward hour by hour to predict further ahead.

Target quantity. Let Y be MAP averaged over the next W hours. The interventional contrast of a patient is the difference in expected outcome with and without a vasopressor,

$$\tau(H_t) \;=\; \mathbb{E}\Big[Y^{\,\mathrm{do}(A=1)} \mid H_t\Big] - \mathbb{E}\Big[Y^{\,\mathrm{do}(A=0)} \mid H_t\Big]\,. \tag{1}$$

The model's own contrast $\hat{\tau}_\theta(H_t)$ is obtained by rolling the model forward twice from the same history, once with the treatment and once without, and taking the difference. As in prior work, we assume that the recorded history contains the factors that drive both the treatment decision and the outcome, so that τ can be identified from data (Assumption 1).

Training objective. Models are trained on the factual objective, which is the likelihood of what was actually observed,

$$\mathcal{L}_{\mathrm{fact}}(\theta) = -\sum_{i,t} \log p_\theta\Big(x_{t+1}^{(i)}, m_{t+1}^{(i)} \;\Big|\; H_t^{(i)}, a_t^{(i)}\Big)\,, \tag{2}$$

where $a_t^{(i)}$ is the treatment that patient i actually received. Each patient is observed under one treatment only, so the model's predictions under the other treatment never enter this

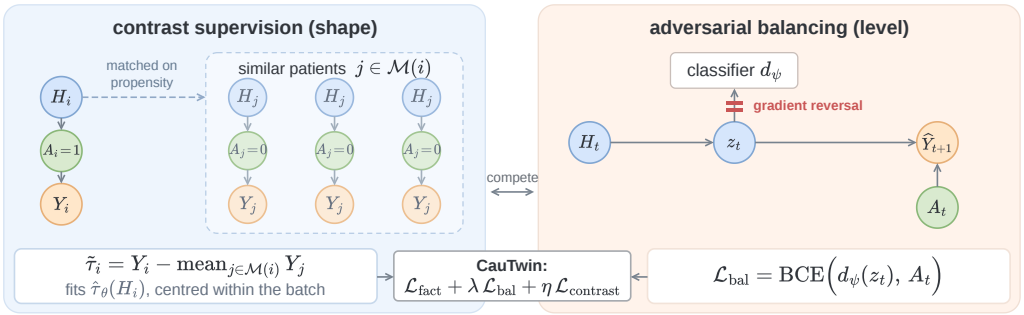


Figure 1: **Overview of CauTwin.** *Left:* contrast supervision compares each patient with similar patients who received the other treatment and sets the shape of the effect. *Right:* adversarial balancing stops the treatment from standing in for severity and sets the level. The two terms compete, and λ and η set the trade-off.

objective. As a result, the objective cannot tell a model with the correct contrast from a model that gives every patient the same contrast, since both fit the data equally well. We prove this in the appendix (Proposition 1).

Consequences. Because training leaves the contrast free, two problems appear in trained models (§5.2). The level is wrong: vasopressors are given when blood pressure is low, and models reproduce this association by predicting that a vasopressor lowers MAP. The shape is missing: the predicted effect is nearly the same for every patient, and replacing a patient's whole history with zeros does not reduce it. A more flexible model does not fix either problem (Appendix D), because the objective still gives it no reason to prefer the correct contrast. What is missing is a training signal for the contrast, and Section 4 provides one.

4 Method

Overview. Section 3 showed that the factual objective leaves both the level and the shape of the contrast unconstrained. CauTwin supplies the missing comparison with similar patients who received the other treatment, uses it only for the shape, and sets the level with adversarial balancing (Figure 1). We describe the prediction model (§4.1), contrast supervision (§4.2), balancing (§4.3), and how the two terms interact (§4.4).

4.1 The prediction model

Backbone. The model is a 10.8M-parameter transformer. Each of the 46 recorded concepts is embedded from its value, whether it was observed, and the hours since it was last observed. The embeddings are pooled per hour. Missing values enter as inputs rather than being filled in, and the model also predicts which concepts will be measured next, because in an ICU the decision to measure is itself informative (Lipton et al., 2016), as also noted in recent EHR imputation work (Dai et al., 2024). We write $z_t = \phi(H_t)$ for the internal state at hour t.

Treatment input. The treatment at hour t enters in two ways. It is added to the input of that hour, and it scales and shifts the internal state before and after the transformer through FiLM conditioning (Perez et al., 2018). An additive input alone would move every patient's prediction by the same amount, so it could not represent an effect that depends on the patient. Section 3 showed that this flexibility is not enough by itself, and contrast supervision is what makes use of it. The output heads predict, for each concept, the mean and variance of its next value and whether it will be measured.

Model contrast. The target we build below averages the outcome over the next W hours, and we define the model's contrast in the same way. From the same history, we roll the model forward twice, once with the treatment on and once with it off, and take the difference:

$$\hat{\tau}_\theta(H_t) = \hat{Y}_\theta^{\,\text{do}(A=1)}(H_t) - \hat{Y}_\theta^{\,\text{do}(A=0)}(H_t), \tag{3}$$

where $\hat{Y}_\theta$ is the model's predicted MAP averaged over the window. Both rollouts follow the predicted mean, so $\hat{\tau}_\theta$ can be trained by gradient descent. The backbone and heads are not our contribution; §5 keeps them fixed and changes only the training objective (Appendix C).

4.2 Counterfactual contrast supervision

Patient matching. Following propensity-score matching (Rosenbaum & Rubin, 1983; Austin, 2011), we fit a model that predicts, at every hour, the probability of receiving the treatment from the record up to that hour. It is trained once on the training split and frozen, and a patient's propensity is its average over the scored window (Figure 1, left). For a treated patient i, its similar patients $\mathcal{M}(i)$ are the k untreated patients with the closest propensity, within a distance δ. For an untreated patient, we search among treated patients in the same way. Patients with fewer than $k/5$ similar patients are left out, because without a comparable patient there is no evidence about the effect. This restricts the contrast loss to histories with empirical treatment overlap (Hess et al., 2026b).

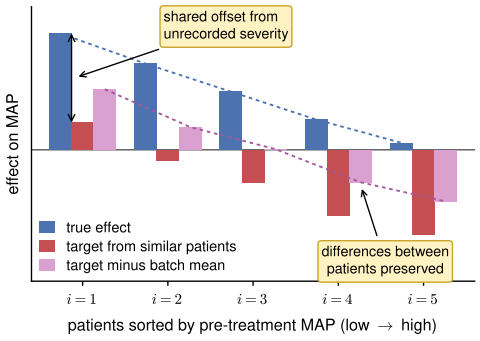


Figure 2: **Why only differences between patients are used.** Targets from similar patients follow the true effect but are shifted as a group by unrecorded severity. Subtracting the batch mean removes the shift and keeps the differences between patients.

Contrast target. For a treated patient, the target is its own outcome minus the average outcome of its similar patients,

$$\tilde{\tau}_i = Y_i - \frac{1}{|\mathcal{M}(i)|} \sum_{j \in \mathcal{M}(i)} Y_j, \tag{4}$$

where Y is MAP averaged over the same W hours. For an untreated patient we reverse the sign, so that a positive target always means that the treatment raised the pressure.

Bias of the target. Similar patients share the chance of being treated, not every clinical detail. Clinicians see severity at the bedside that the records capture only partly, so treated patients tend to be sicker than their similar untreated patients. This shifts all targets downward: on the training split their average is -5.28 mmHg, which would say that a vasopressor lowers blood pressure. The shift, however, moves the targets together. What remains reliable is how they differ between patients, that is, which patients respond more and which respond less (Figure 2).

Contrast loss. We subtract the batch average from both the model's contrast and the target, and fit only what is left:

$$\mathcal{L}_{\text{contrast}} = \frac{1}{|B|} \sum_{i \in B} \ell_\kappa \left(\frac{\hat{\tau}_\theta(H_i) - \overline{\hat{\tau}_\theta}}{s}, \; \frac{\tilde{\tau}_i - \overline{\tilde{\tau}}}{s} \right), \tag{5}$$

where B is the batch, ℓ_κ is the Huber loss, which limits the influence of a few noisy targets, and s is the standard deviation of the centred targets, which puts this term on the same scale as the likelihood.

Theoretical guarantee. Subtracting the average removes any error shared by all patients, so we never need to estimate that error. If similar patients differ from the patient mainly by an offset that is the same for everyone, the loss recovers how the effect varies across patients exactly; if the offset varies, the error is no larger than that variation (Assumption 2 and Proposition 2 in Appendix A.3). The price is that the loss says nothing about the average effect, which must come from elsewhere.

4.3 Adversarial balancing for the level

Source of the level bias. Clinicians give a vasopressor to patients who are sicker, so the treatment input itself tells the model that the patient is severe. The model learns "treated, so severe, so low pressure", so a rollout with the treatment switched on predicts a fall.

Adversarial balancing. We block this shortcut with the balancing term of Bica et al. (2020b), used unchanged. A small classifier d_ψ tries to tell from the internal state z_t whether the patient is being treated (Figure 1, right):

$$\mathcal{L}_{\text{bal}} = \text{BCE}\big(d_\psi(z_t),\, A_t\big). \tag{6}$$

The classifier is trained to succeed and the model is trained to make it fail, through gradient reversal (Ganin et al., 2016). The treatment input can then no longer stand in for severity, and the effect head has to describe what the treatment itself does. Balancing alone cannot rule out unrecorded confounding (Bica et al., 2020a; Schweisthal et al., 2024). Because it works on the recorded data, we treat the level it produces as anchored rather than identified, and verify its direction against pharmacology in §5. Two training rules, applied to every model, matter: λ is increased gradually, and no checkpoint is selected while it increases (Appendix C).

4.4 Training objective

Objective. The full training objective is

$$\mathcal{L}(\theta) = \underbrace{\mathcal{L}_{\text{fact}}}_{\text{trajectory}} + \lambda \underbrace{\mathcal{L}_{\text{bal}}}_{\text{level}} + \eta \underbrace{\mathcal{L}_{\text{contrast}}}_{\text{shape}} + \mu \underbrace{\mathcal{L}_{\text{phys}}}_{\text{physiology}}. \tag{7}$$

Here λ, η and μ weight the three added terms.

Physiological consistency. The last term penalises violations of two known relations in the model's predictions, the Henderson-Hasselbalch equation and the ordering SBP $\geq$ MAP $\geq$ DBP, wherever the values involved are measured. It keeps predicted states physically possible, also under treatments not given, and sets neither the level nor the shape.

Interaction of the two terms. Balancing and contrast supervision act on the same model in opposite directions. Balancing requires that the treatment pathway not depend on severity, so that the treatment cannot stand in for it. Contrast supervision requires that the size of the effect depend on severity, since a patient in shock should respond more than a patient with normal pressure. Balancing alone therefore tends to flatten the effect across patients, and contrast supervision alone leaves the average effect at the confounded value. The weights λ and η set the balance between the two. Letting balancing act only on the part of the model that produces the effect does not remove this tension, because confounding then reaches the effect through the unbalanced part (§5.3, Appendix D.2). We therefore apply both terms to the same model.

5 Experiments

5.1 Setup

ICU data. We use three ICU cohorts from separate US hospital systems: NWICU, eICU-CRD and MIMIC-IV (Johnson et al., 2023; Pollard et al., 2018). All three are mapped to a shared set of 46 clinical concepts, and we keep stays with at least one MAP and one lactate measurement in the first 24 hours. This gives 47,674 trajectories of 48 hours. The treatment of interest is a vasopressor, and the outcome is MAP over the following 12 hours. Data are split by stay, and the cohorts are described in Appendix B.

Semi-synthetic data. Real data contain no counterfactual outcomes, so we also use the tumour growth simulator of Bica et al. (2020b), where $\gamma \in \{1, 2, 3, 4\}$ sets the strength of confounding and counterfactual error can be measured directly. Treatment there depends only on the recorded tumour state, so this benchmark tests the prediction model and its

Table 1: Vasopressor effect on MAP for ICU test patients. Correct direction: share of patients for whom the predicted effect is positive. Patient dependence: rank correlation between predicted effect and pre-treatment MAP, which should be negative. Mean ± 95% CI over seeds; †single seed.

Model	Effect (mmHg)	Correct direction (%) ↑	Patient dependence ρ ↓
GRU	-2.73 ± 0.30	29.5 ± 1.8	-0.19 ± 0.04
RSSM	-4.57 ± 1.23	28.0 ± 6.0	-0.17 ± 0.22
CRN ($\lambda = 3$)	-1.40 ± 1.94	40.0 ± 15.8	-0.10 ± 0.06
CRN ($\lambda = 0.3$, its best)†	$+0.50$	55.8	-0.12
IPW	-3.44 ± 0.68	29.5 ± 6.7	-0.19 ± 0.20
G-computation	-2.82 ± 1.24	33.6 ± 11.5	-0.18 ± 0.09
CauTwin (ours)	$+6.90 \pm 1.19$	$\mathbf{79.7 \pm 1.1}$	$\mathbf{-0.46 \pm 0.10}$

Table 2: Counterfactual RMSE (↓) on the tumour growth benchmark; γ is the strength of confounding. Best in **bold**, second underlined; mean over 3 seeds. Four further baselines and standard deviations are in Appendix E.3.

Method	γ=1	γ=2	γ=3	γ=4	Avg
GRU	0.488	0.412	0.404	0.398	0.426
CRN	0.486	0.418	0.395	0.363	0.416
RMSN	0.433	0.398	0.392	0.381	0.401
Causal Transformer	0.465	0.392	0.366	0.343	0.392
WO-Learner	0.555	0.466	0.447	0.443	0.478
G-Net	0.445	0.395	0.368	0.366	0.394
IGC-Net	0.455	0.373	0.353	0.334	0.379
CauTwin (ours)	**0.434**	**0.364**	**0.330**	**0.329**	**0.364**

counterfactual rollouts, while contrast supervision, which targets severity missing from the records, is tested on ICU data (Appendix E).

Evaluation metrics. Without counterfactual outcomes, we check the two properties of the contrast against what pharmacology requires. *Correct direction* is the share of patients for whom the model predicts that a vasopressor raises MAP, which checks the level. *Patient dependence* is the rank correlation ρ between the predicted effect and the patient's MAP before treatment, which checks the shape. A patient with lower pressure should gain more, so ρ should be negative, and a value near zero means that the effect ignores the patient (Appendix C.5).

Baselines and protocol. We compare with a GRU and a recurrent state-space model (RSSM) (Hafner et al., 2020) trained on the factual objective, CRN (Bica et al., 2020b), and two adjustment strategies built on our backbone: inverse propensity weighting as in RMSN (Lim et al., 2018), and G-computation as in IGC-Net (Hess et al., 2026a). All ICU models have 10.8M parameters (±9%) and follow the same training and checkpoint rules. We report the mean and a 95% confidence interval over 3 to 5 seeds, on 450 patients drawn from the test split, 150 from each hospital. On the tumour benchmark we compare with eleven methods at their published settings (Appendix C).

5.2 Main results

Direction of the effect. Table 1 and Figure 3a show that all methods trained on the factual objective predict that a vasopressor lowers MAP for most patients. This includes the methods that adjust for confounding. CRN reaches at most 56% at its best setting, and inverse propensity weighting and G-computation stay near 30%. This is the level failure of Section 3: the models reproduce the association between treatment and low pressure. CauTwin predicts a rise in MAP for 80% of patients.

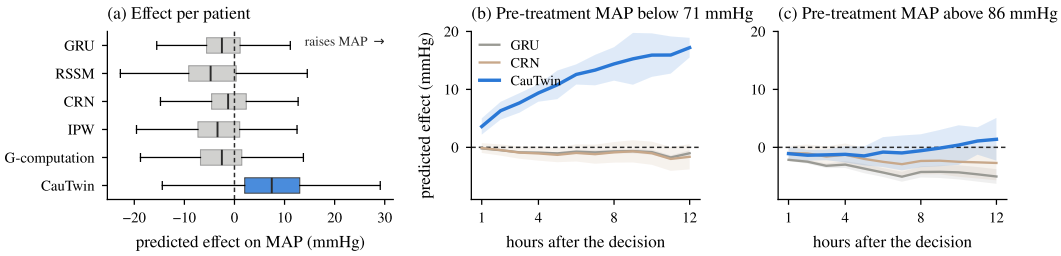


Figure 3: **Predicted vasopressor effects on ICU test patients.** (a) Effect per patient, pooled over seeds. Every baseline places most patients below zero. (b, c) Effect over the 12 hours after the decision, for patients with low and with normal pre-treatment MAP. Bands show 95% confidence intervals over seeds.

Table 3: Ablation of the two training terms on ICU data. λ: weight of balancing; η: weight of contrast supervision. The last two rows try to separate the two terms inside the model. Mean ± 95% CI over 3 to 5 seeds.

Variant	λ	η	Correct direction (%) ↑	Patient dependence ρ ↓
Factual objective only	0	0	30.7 ± 7.1	-0.27 ± 0.08
Balancing only	3	0	86.4 ± 21.2	$+0.06 \pm 0.13$
Contrast supervision only	0	1	38.2 ± 2.5	-0.58 ± 0.07
CauTwin	3	1	79.7 ± 1.1	-0.46 ± 0.10
Stronger contrast supervision	3	3	70.6 ± 27.6	-0.53 ± 0.11
Balancing on part of the state	3	1	33.8 ± 11.0	-0.55 ± 0.03
Treatment as residual, no classifier	0	1	39.5 ± 1.9	-0.49 ± 0.06

Dependence on the patient. The baselines show a weak negative correlation with pre-treatment MAP, but they attach it to an effect of the wrong direction. CauTwin gives an effect of the correct direction whose size clearly depends on the patient ($\rho = -0.46$). For patients with low pressure, it predicts a rise that builds up over the 12 hours, while GRU and CRN predict a fall (Figure 3b). For patients with normal pressure, it predicts almost no change (Figure 3c). It is the only model with both properties.

Tumour benchmark. In Table 2, CauTwin has the lowest error at every confounding level, 4.0% below the second-best method on average (Appendix E.3).

5.3 Ablation

Role of each term. Table 3 and Figure 4 separate the two terms. Balancing alone fixes the direction but removes any dependence on the patient ($\rho \approx 0$), so every quartile receives a similar effect. Contrast supervision alone produces the right ordering across quartiles but leaves the direction wrong. Only the two together give both.

Trade-off between the terms. Adding contrast supervision to balancing lowers the correct direction slightly, and raising its weight from $\eta = 1$ to $\eta = 3$ lowers it further while the patient dependence barely improves. This is the trade-off described in §4.4. We use $\eta = 1$, where the patient dependence has already saturated.

Separating the two terms. The last two rows keep severity away from the treatment without balancing the whole model: balancing acts only on the part of the state that the effect reads, or the treatment enters as its difference from the predicted propensity. Both learn the shape but lose the direction (Appendix D.2).

Physiological term. Without it, the balanced model keeps the correct direction (82.9%, single seed). With it, violations of SBP ≥ MAP ≥ DBP under treatments not given fall from 1.27% to 0.02% (Appendix D).

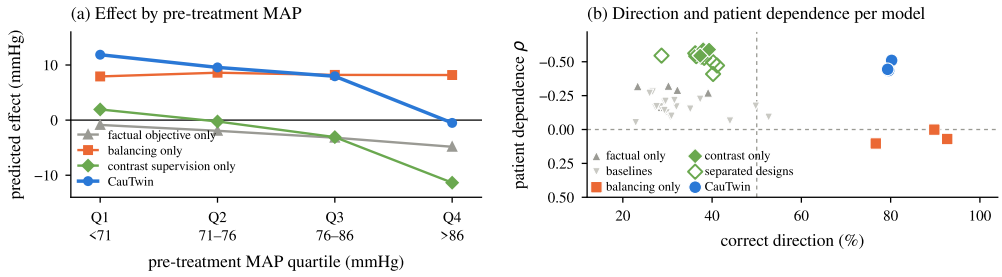


Figure 4: **The two training terms on ICU test patients.** (a) Predicted effect by quartile of pre-treatment MAP for the four configurations of Table 3, mean over seeds. Balancing alone gives every patient a similar positive effect. Contrast supervision alone orders patients correctly but at a negative level. CauTwin combines both. (b) One point per trained model; the axes check the level and the shape. Only models with both terms reach the upper right.

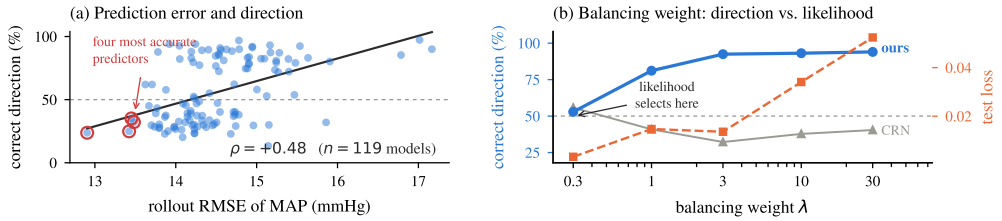


Figure 5: (a) Each point is one of 119 trained models: better prediction of the observed future goes with a worse direction of the effect. (b) Along the balancing sweep, likelihood and correct direction move in opposite directions.

5.4 Clinical validity and model selection

Clinical plausibility. Patients in the lowest quartile of pre-treatment MAP, below 71 mmHg, receive a predicted rise of 11.9 mmHg, while patients above 86 mmHg receive almost none (Figures 3b,c and 4a). Without retuning, the correction carries over to systolic (86.4% correct direction) and diastolic pressure (78.4%), and it holds in all three hospitals (73.1% to 89.8%). Every seed also predicts the baroreflex fall in heart rate after a vasopressor, although no term in the objective encodes it.

Cost in factual accuracy. Balancing removes a shortcut that helps predict the observed future, so its cost should appear only where a vasopressor was given. It does: MAP rollout error rises from 13.6 to 15.9 mmHg on treated patients and stays at 14.4 to 14.5 mmHg on untreated ones (five seeds, Appendix D).

Model selection. Missing counterfactual outcomes complicate model selection (Curth & van der Schaar, 2023; Gao, 2025). Across 119 trained models, a lower rollout error goes with a worse direction ($\rho = +0.48$, Figure 5a). Choosing the balancing weight by validation likelihood selects $\lambda = 0$, the causally worst setting at 31% correct direction, while larger weights reach 94% (Figure 5b).

6 Conclusion

Training on observed outcomes guarantees neither that a predicted treatment effect points in the right direction nor that it differs between patients. CauTwin supplies the missing signal by comparing each patient with similar patients who received the other treatment, learns only how the effect differs between patients, and leaves the average effect to adversarial balancing. On three ICU cohorts, it is the only model whose vasopressor effects both point in the correct direction and depend on the patient. Next steps include doses, treatment timing and prospective validation of effect sizes (Appendix F).

AI use statement

We used generative AI tools to assist with language editing, code development, figure preparation, and the presentation of experimental analyses. In addition, a language model is one component of our data harmonisation pipeline, where it proposes mappings from database schemas to shared concepts using schema information only, and every proposal is checked against the source dictionary (Appendix B.2). The authors reviewed all AI-assisted content, tested the AI-assisted code, and checked the reported analyses against experimental records. We take responsibility for the final text, results, claims, and artifacts.

Ethics statement

This work uses three de-identified ICU databases, MIMIC-IV, eICU-CRD and NWICU, which were collected and de-identified by their source institutions and are released for research through PhysioNet under credentialed access. We accessed them under their data use agreements, made no attempt to re-identify patients, and do not redistribute any patient data. No patient data were sent to any language model. Our models predict the effect of treatments that were not given, and like any such model they require prospective validation before their predictions inform care. Evaluating the models across demographic subgroups and in health systems outside the United States is an important next step.

Reproducibility statement

Appendix A states all assumptions and gives complete proofs of the theoretical results. Appendix B describes cohort selection, harmonisation, the concept list, the treatment definition and the data splits, and Appendix C describes the model architecture, the matching procedure, all hyperparameters and the baselines. Results for every seed are reported in Appendix D, and the tumour simulator and its evaluation protocol are specified in Appendix E. All models are evaluated on the same fixed split and the same test patients, and confidence intervals are defined in Appendix C.5. The code for harmonisation, training and evaluation, together with the tumour simulator, is available at `https://anonymous.4open.science/r/CauTwin-72F7/`. The ICU databases are available on PhysioNet to credentialed users.

References

Peter C. Austin. Optimal caliper widths for propensity-score matching when estimating differences in means and differences in proportions in observational studies. *Pharmaceutical Statistics*, 10(2):150–161, 2011. doi: 10.1002/pst.433.

Sébastien Benzekry, Clare Lamont, Afshin Beheshti, et al. Classical mathematical models for description and prediction of experimental tumor growth. *PLoS Computational Biology*, 10(8), 2014.

Ioana Bica, Ahmed Alaa, and Mihaela van der Schaar. Time series deconfounder: Estimating treatment effects over time in the presence of hidden confounders. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 884–895, 2020a. URL `https://proceedings.mlr.press/v119/bica20a.html`.

Ioana Bica, Ahmed M Alaa, James Jordon, and Mihaela van der Schaar. Estimating counterfactual treatment outcomes over time through adversarially balanced representations. In *ICLR*, 2020b.

Isabel Chien, Cliff Wong, Zelalem Gero, Jaspreet Bagga, Risa Ueno, Richard E. Turner, Roshanthi K. Weerasinghe, Brian Piening, Tristan Naumann, Carlo Bifulco, Hoifung Poon, and Javier González Hernández. Beyond clinical trials: Using real world evidence to investigate heterogeneous, time-varying treatment effects. In *Proceedings of the 9th Machine Learning for Healthcare Conference*, volume 252 of *Proceedings of Machine Learning Research*, 2024. URL `https://proceedings.mlr.press/v252/chien24a.html`.

Yoichi Chikahara and Kansei Ushiyama. Differentiable pareto-smoothed weighting for high-dimensional heterogeneous treatment effect estimation. In *Proceedings of the Fortieth Conference on Uncertainty in Artificial Intelligence*, volume 244 of *Proceedings of Machine Learning Research*, pp. 749–762, 2024. URL https://proceedings.mlr.press/v244/chikahara24a.html.

Alicia Curth and Mihaela van der Schaar. In search of insights, not magic bullets: Towards demystification of the model selection dilemma in heterogeneous treatment effect estimation. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 6623–6642, 2023. URL https://proceedings.mlr.press/v202/curth23b.html.

Zongyu Dai, Emily Getzen, and Qi Long. SADI: Similarity-aware diffusion model-based imputation for incomplete temporal EHR data. In *Proceedings of The 27th International Conference on Artificial Intelligence and Statistics*, volume 238 of *Proceedings of Machine Learning Research*, pp. 4195–4203, 2024. URL https://proceedings.mlr.press/v238/dai24c.html.

Leon Deng, Hong Xiong, Feng Wu, Sanyam Kapoor, Soumya Gosh, Zach Shahn, and Li-wei Lehman. Uncertainty quantification for conditional treatment effect estimation under dynamic treatment regimes. In *Proceedings of the 4th Machine Learning for Health Symposium*, volume 259 of *Proceedings of Machine Learning Research*, pp. 248–266, 2025. URL https://proceedings.mlr.press/v259/deng25a.html.

Mouad El Bouchattaoui, Myriam Tami, Benoit Lepetit, and Paul-Henry Cournède. Causal contrastive learning for counterfactual regression over time. In *Advances in Neural Information Processing Systems*, volume 37, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/02cef2ae63853724eb99e70721d3bc65-Abstract.html.

Aaron Fisher. Inverse-variance weighting for estimation of heterogeneous treatment effects. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 13706–13733, 2024. URL https://proceedings.mlr.press/v235/fisher24a.html.

Dennis Frauen, Konstantin Hess, and Stefan Feuerriegel. Model-agnostic meta-learners for estimating heterogeneous treatment effects over time. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/d51ab0fc62fe2d777c7569952f518f56-Abstract-Conference.html.

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, et al. Domain-adversarial training of neural networks. *JMLR*, 17, 2016.

Zijun Gao. Trustworthy assessment of heterogeneous treatment effect estimator via analysis of relative error. In *Proceedings of The 28th International Conference on Artificial Intelligence and Statistics*, volume 258 of *Proceedings of Machine Learning Research*, pp. 1063–1071, 2025. URL https://proceedings.mlr.press/v258/gao25d.html.

Omer Gottesman, Fredrik Johansson, Matthieu Komorowski, et al. Guidelines for reinforcement learning in healthcare. *Nature Medicine*, 25, 2019.

Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. In *Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=tEYskw1VY2.

Danijar Hafner, Timothy Lillicrap, Jimmy Ba, and Mohammad Norouzi. Dream to control: Learning behaviors by latent imagination. In *ICLR*, 2020.

Miguel A. Hernán and James M. Robins. *Causal Inference: What If*. Chapman & Hall/CRC, 2020.

Konstantin Hess, Dennis Frauen, Valentyn Melnychuk, and Stefan Feuerriegel. IGC-Net for conditional average potential outcome estimation over time. In *ICLR*, 2026a.

Konstantin Hess, Dennis Frauen, Mihaela van der Schaar, and Stefan Feuerriegel. Overlap-weighted orthogonal meta-learner for treatment effect estimation over time. In *ICLR*, 2026b.

Qiang Huang, Chuizheng Meng, Defu Cao, Biwei Huang, Yi Chang, and Yan Liu. An empirical examination of balancing strategy for counterfactual estimation on time series. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 20043–20062, 2024. URL https://proceedings.mlr.press/v235/huang24r.html.

Emil Javurek, Valentyn Melnychuk, Jonas Schweisthal, Konstantin Hess, Dennis Frauen, and Stefan Feuerriegel. An orthogonal learner for individualized outcomes in Markov decision processes. In *ICLR*, 2026.

Russell Jeter, Christopher Josef, Supreeth Shashikumar, and Shamim Nemati. Does the artificial intelligence clinician learn optimal treatment strategies for sepsis in intensive care? *arXiv:1902.03271*, 2019.

Fredrik Johansson, Uri Shalit, and David Sontag. Learning representations for counterfactual inference. In *ICML*, 2016.

Alistair E W Johnson, Lucas Bulgarelli, Lu Shen, et al. Mimic-iv, a freely accessible electronic health record dataset. *Scientific Data*, 10, 2023.

Rickard Karlsson and Jh Krijthe. Falsification of unconfoundedness by testing independence of causal mechanisms. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 29128–29147, 2025. URL https://proceedings.mlr.press/v267/karlsson25a.html.

Edward H. Kennedy. Towards optimal doubly robust estimation of heterogeneous causal effects. *Electronic Journal of Statistics*, 17(2), 2023. doi: 10.1214/23-EJS2157.

Patrick Kidger, James Morrill, James Foster, and Terry Lyons. Neural controlled differential equations for irregular time series. *NeurIPS*, 2020.

Matthieu Komorowski, Leo A Celi, Omar Badawi, et al. The artificial intelligence clinician learns optimal treatment strategies for sepsis in intensive care. *Nature Medicine*, 24, 2018.

Rui Li, Stephanie Hu, Mingyu Lu, Yuria Utsumi, Prithwish Chakraborty, Daby M. Sow, Piyush Madan, Jun Li, Mohamed Ghalwash, Zach Shahn, and Li-wei Lehman. G-Net: a recurrent network approach to g-computation for counterfactual prediction under a dynamic treatment regime. In *Proceedings of Machine Learning for Health*, volume 158 of *Proceedings of Machine Learning Research*, pp. 282–299, 2021. URL https://proceedings.mlr.press/v158/li21a.html.

Bryan Lim, Ahmed Alaa, and Mihaela van der Schaar. Forecasting treatment responses over time using recurrent marginal structural networks. In *NeurIPS*, 2018.

Zachary C. Lipton, David Kale, and Randall Wetzel. Directly modeling missing data in sequences with RNNs: Improved classification of clinical time series. In *Proceedings of the 1st Machine Learning for Healthcare Conference*, volume 56 of *Proceedings of Machine Learning Research*, pp. 253–270, 2016. URL https://proceedings.mlr.press/v56/Lipton16.html.

Valentyn Melnychuk, Dennis Frauen, and Stefan Feuerriegel. Causal transformer for estimating counterfactual outcomes. In *ICML*, 2022.

Valentyn Melnychuk, Dennis Frauen, and Stefan Feuerriegel. Bounds on representation-induced confounding bias for treatment effect estimation. In *ICLR*, 2024.

Xinkun Nie and Stefan Wager. Quasi-oracle estimation of heterogeneous treatment effects. *Biometrika*, 108, 2021.

Ethan Perez, Florian Strub, Harm De Vries, Vincent Dumoulin, and Aaron Courville. Film: Visual reasoning with a general conditioning layer. In *AAAI*, 2018.

Tom J Pollard, Alistair E W Johnson, Jesse D Raffa, et al. The eicu collaborative research database. *Scientific Data*, 5, 2018.

James Robins. A new approach to causal inference in mortality studies. *Mathematical Modelling*, 7, 1986.

James M Robins, Miguel A Hernán, and Babette Brumback. Marginal structural models and causal inference in epidemiology. *Epidemiology*, 11, 2000.

Paul R Rosenbaum and Donald B Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70, 1983.

Jonas Schweisthal, Dennis Frauen, Mihaela van der Schaar, and Stefan Feuerriegel. Meta-learners for partially-identified treatment effects across multiple environments. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 43967–43985, 2024. URL https://proceedings.mlr.press/v235/schweisthal24a.html.

Jonas Schweisthal, Dennis Frauen, Maresa Schröder, Konstantin Hess, Niki Kilbertus, and Stefan Feuerriegel. Learning representations of instruments for partial identification of treatment effects. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 53703–53727, 2025. URL https://proceedings.mlr.press/v267/schweisthal25a.html.

Nabeel Seedat, Fergus Imrie, Alexis Bellot, Zhaozhi Qian, and Mihaela van der Schaar. Continuous-time modeling of counterfactual outcomes using neural controlled differential equations. In *ICML*, 2022.

Uri Shalit, Fredrik D Johansson, and David Sontag. Estimating individual treatment effect: Generalization bounds and algorithms. In *ICML*, 2017.

Toru Shirakawa, Yi Li, Yulun Wu, Sky Qiu, Yuxuan Li, Mingduo Zhao, Hiroyasu Iso, and Mark J. van der Laan. Longitudinal targeted minimum loss-based estimation with temporal-difference heterogeneous transformer. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 45097–45113, 2024. URL https://proceedings.mlr.press/v235/shirakawa24a.html.

Shengpu Tang and Jenna Wiens. Model-based reinforcement learning for offline zero-shot planning. *Proceedings of Machine Learning Research*, 2022.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding, 2018. URL https://arxiv.org/abs/1807.03748.

Lu Wang, Wei Zhang, Xiaofeng He, and Hongyuan Zha. Supervised reinforcement learning with recurrent neural network for dynamic treatment recommendation. *KDD*, 2018.

Xin Wang, Shengfei Lyu, Lishan Yang, Yibing Zhan, and Huanhuan Chen. A dual-module framework for counterfactual estimation over time. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 51063–51085, 2024. URL https://proceedings.mlr.press/v235/wang24au.html.

Xin Wang, Shengfei Lyu, Chi Luo, Xiren Zhou, and Huanhuan Chen. Variational counterfactual intervention planning to achieve target outcomes. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 63860–63881, 2025. URL https://proceedings.mlr.press/v267/wang25bv.html.

Jiaxuan Zhang, Emadeldeen Eldele, Fuyuan Cao, Yang Wang, Xiaoli Li, and Jiye Liang. Counterfactual contrastive learning with normalizing flows for robust treatment effect estimation. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 74819–74839, 2025. URL https://proceedings.mlr.press/v267/zhang25r.html.

Appendix

A Proofs **15**
A.1 Assumptions . . . 15
A.2 Proposition 1: Observed Outcomes Do Not Determine the Effect . . . 15
A.3 Proposition 2: Contrast Supervision Recovers Differences Between Patients . 16

B Data **17**
B.1 Cohorts . . . 17
B.2 Harmonisation . . . 18
B.3 Treatment . . . 19

C Implementation **19**
C.1 Model . . . 19
C.2 Matching Similar Patients . . . 19
C.3 Training and Hyperparameters . . . 20
C.4 Baselines . . . 21
C.5 Evaluation Metrics . . . 21

D More ICU Results **22**
D.1 Ablation . . . 22
D.2 Balancing . . . 23
D.3 Generalisation . . . 24
D.4 Model Selection . . . 24

E Tumour Benchmark **25**
E.1 Setup . . . 25
E.2 Model . . . 26
E.3 Results . . . 26

F Scope and Future Work **28**

G Related Work **28**

A Proofs

This section states the assumptions behind our results and proves the two propositions used in the main text. Throughout, for a history h and a treatment $a \in \{0,1\}$, we write $\mu_a(h) = \mathbb{E}[Y^{\mathrm{do}(A=a)} \mid H = h]$ for the expected outcome under treatment a, so that $\tau(h) = \mu_1(h) - \mu_0(h)$. For a model with parameters θ, $\hat{\mu}_a^\theta(h)$ is its predicted outcome under treatment a, obtained by rollout, and $\hat{\tau}_\theta(h) = \hat{\mu}_1^\theta(h) - \hat{\mu}_0^\theta(h)$ is its contrast.

A.1 Assumptions

Our first assumption is the standard condition under which treatment effects can be identified from observational data.

Assumption 1 (Sequential ignorability)**.** *For every hour t and $a \in \{0,1\}$: (i) consistency, $Y = Y^{\mathrm{do}(A=A_t)}$; (ii) no unmeasured confounding, $A_t \perp\!\!\!\perp Y^{\mathrm{do}(A=a)} \mid H_t$; (iii) positivity, $0 < \mathbb{P}(A_t = 1 \mid H_t) < 1$ almost surely.*

Under Assumption 1, the expected outcome under each treatment equals an observable conditional mean:

$$\mu_a(h) = \mathbb{E}[Y^{\mathrm{do}(A=a)} \mid H = h] = \mathbb{E}[Y^{\mathrm{do}(A=a)} \mid H = h, A = a] = \mathbb{E}[Y \mid H = h, A = a], \quad (8)$$

where the second step uses (ii) and (iii) and the third uses (i).

Our second assumption concerns the targets built from similar patients. We call the systematic part of their error the matching error,

$$b(h) = \mathbb{E}\left[\tilde{\tau}_i - \tau(H_i) \mid H_i = h\right], \quad (9)$$

and write $\varepsilon_i = \tilde{\tau}_i - \tau(H_i) - b(H_i)$ for the remaining noise, so that $\tilde{\tau}_i = \tau(H_i) + b(H_i) + \varepsilon_i$ with $\mathbb{E}[\varepsilon_i \mid H_i] = 0$. For a treated patient, the matching error is the difference between how the patient and its similar patients would fare without treatment. It is zero when the similar patients are fully comparable, and it is negative in our data because treated patients are sicker than their matches in ways the records do not capture.

Assumption 2 (Similar patients are comparable up to a shared offset)**.** *The matching error is the same for all patients: $b(h) = b$ for a constant b.*

This shared-offset assumption is stronger than conventional propensity-score matching and is not implied by sequential ignorability. With unmeasured confounding or weak overlap, point identification may fail (Schweisthal et al., 2024; Karlsson & Krijthe, 2025).

A.2 Proposition 1: Observed Outcomes Do Not Determine the Effect

Let $\mathcal{D} = \{(h_i, a_i, y_i)\}_{i=1}^n$ be the training data and $\hat{\mathcal{L}}(\theta) = -\frac{1}{n}\sum_i \log p_\theta(y_i \mid h_i, a_i)$ the factual objective. We call a model class Θ rich if, starting from any $\theta_0 \in \Theta$, it can change the predictions for the treatments that were not given without changing the predictions for the observed pairs (h_i, a_i). Large neural networks have this property on finite data.

Proposition 1 (Contrast degeneracy)**.** *Suppose the training histories are pairwise distinct and Θ is rich. Let θ_0 minimise $\hat{\mathcal{L}}$. Then for every function u of the history there is a minimiser θ_u with $\hat{\mathcal{L}}(\theta_u) = \hat{\mathcal{L}}(\theta_0)$ and $\hat{\tau}_{\theta_u}(h_i) = u(h_i)$ for all i.*

Proof. For each patient, choose the prediction under the treatment that was not given as

$$m_i = \begin{cases} \hat{\mu}_1^{\theta_0}(h_i) - u(h_i) & \text{if } a_i = 1, \\ \hat{\mu}_0^{\theta_0}(h_i) + u(h_i) & \text{if } a_i = 0. \end{cases} \quad (10)$$

Since the histories are distinct, no pair $(h_i, 1 - a_i)$ coincides with an observed pair, so these choices do not conflict with the data. Richness gives a θ_u that agrees with θ_0 on every observed pair and predicts m_i on every unobserved pair. The objective depends only on the observed pairs, so $\hat{\mathcal{L}}(\theta_u) = \hat{\mathcal{L}}(\theta_0)$, and by construction $\hat{\tau}_{\theta_u}(h_i) = u(h_i)$.

For a model that predicts hour by hour, the argument applies at every step: a counterfactual rollout queries the model at the same history with the other treatment and then at histories it has generated itself, and none of these queries enters the objective. Enlarging the model class does not change this conclusion.

Corollary 1 (A more flexible model does not help)**.** *If $\Theta \subseteq \Theta'$ and Θ is rich, then the minimisers of $\hat{\mathcal{L}}$ over Θ' also realise every contrast on the training histories.*

This matches Appendix D.1, where additive and FiLM treatment inputs give indistinguishable effects. With unlimited data and overlap, the factual objective would recover $\mathbb{E}[Y \mid H = h, A = a]$ for every pair with positive density, and (8) would then identify τ. In high-dimensional ICU histories, however, each history is seen under one treatment only, so the prediction for the other treatment is set by how the model generalises, which is what Proposition 1 formalises.

A.3 Proposition 2: Contrast Supervision Recovers Differences Between Patients

We analyse (5) with the squared loss and a fixed scale s. Writing $d_i = \hat{\tau}_\theta(H_i) - \tilde{\tau}_i$, both arguments of the loss are centred with the same batch mean, so the batch loss is a sample variance, and its expectation over batches of independent patients is

$$\mathbb{E}\left[\ell_B(\theta)\right] = \mathbb{E}\left[\frac{1}{|B|\, s^2} \sum_{i \in B} \big(d_i - \bar{d}\big)^2\right] = \frac{|B| - 1}{|B|\, s^2} \operatorname{Var}(d). \tag{11}$$

Training therefore minimises the variance of d across patients.

Proposition 2 (Shape identification)**.** *Any minimiser of $\mathbb{E}[\ell_B(\theta)]$ over an unrestricted $\hat{\tau}_\theta$ satisfies, almost surely,*

$$\hat{\tau}_\theta(H) - \mathbb{E}[\hat{\tau}_\theta] = \big(\tau(H) - \mathbb{E}[\tau]\big) + \big(b(H) - \mathbb{E}[b]\big), \tag{12}$$

and the objective does not change when a constant is added to $\hat{\tau}_\theta$. Under Assumption 2 the shape of the contrast is recovered exactly; otherwise its error is the variation of the matching error across patients.

Proof. Write $d = g(H) - \varepsilon$ with $g(H) = \hat{\tau}_\theta(H) - \tau(H) - b(H)$. Because $\mathbb{E}[\varepsilon \mid H] = 0$, the two parts are uncorrelated and

$$\operatorname{Var}(d) = \operatorname{Var}\big(g(H)\big) + \operatorname{Var}(\varepsilon). \tag{13}$$

The second term does not depend on θ, and the first is zero exactly when $g(H)$ is constant, which is (12). Adding a constant to $\hat{\tau}_\theta$ shifts every d_i equally and leaves $d_i - \bar{d}$ unchanged.

The Huber loss used in practice equals the squared loss for small residuals and limits the influence of a few noisy targets. Without centering, the same argument gives $\hat{\tau}_\theta = \tau + b$, so the level would inherit the matching error of -5.28 mmHg. Combining the two propositions gives the role of each training term.

Corollary 2 (What each term controls)**.** *Under Assumption 2, the contrast term determines the shape $\hat{\tau}_\theta - \mathbb{E}[\hat{\tau}_\theta]$ and places no constraint on the level $\mathbb{E}[\hat{\tau}_\theta]$. Without the contrast term, the shape is free as in Proposition 1. Without the balancing term, the level is set by the factual objective alone.*

Sensitivity to a varying matching error. Proposition 2 also covers the case where Assumption 2 fails: the learned shape is the true shape plus the centred matching error $b(H) - \mathbb{E}[b]$. The error in the shape therefore has variance $\operatorname{Var}\big(b(H)\big)$, and the ordering of patients is kept as long as the matching error varies less than the effect itself. Three checks bear on this in our data.

- **Hospitals.** A matching error that differs between hospitals would appear as a difference between hospitals, not within them. The patient dependence of CauTwin is negative within every hospital and every seed, with mean ρ of -0.49 in eICU, -0.31 in MIMIC-IV and -0.58 in NWICU.

- **Severity.** For the matching error alone to produce the negative dependence we find, it would have to be higher for patients with lower pressure, meaning that treated patients with low pressure differ less in severity from their matches than other treated patients do. If the severity gap between treated patients and their matches is instead largest among patients with low pressure, as confounding by indication suggests, the matching error lowers the learned effect for these patients and weakens the dependence rather than creating it.

- **Treatment course.** The targets reflect the course of treatment that was given (Appendix C.2). Its length is almost unrelated to pre-treatment MAP (Spearman -0.08 among treated training stays), so differences in the course cannot account for the dependence on pre-treatment pressure.

Proof. The first claim is Proposition 2. The balancing loss depends only on the internal state and the observed treatment, so Proposition 1 applies unchanged to $\hat{\mathcal{L}} + \lambda \mathcal{L}_{\text{bal}}$, which gives the second claim. The third holds because the contrast loss is unchanged by a constant shift of $\hat{\tau}_\theta$.

B Data

This section describes the three ICU cohorts, how they are mapped to a shared schema, and how treatments are defined.

B.1 Cohorts

Inclusion. We use three ICU databases from separate US hospital systems: NWICU, eICU-CRD (Pollard et al., 2018) and MIMIC-IV (Johnson et al., 2023). We keep the first ICU stay of each hospital admission for adults with a stay of at least six hours, and require at least one MAP and one lactate measurement in the first 24 hours. Each remaining stay contributes its first 48 hours. Table 4 lists every screening step.

Table 4: Screening. Each row gives the number of ICU stays that remain after the step.

Step	NWICU	eICU	MIMIC-IV
ICU stays, raw	28,404	198,875	94,266
Valid admission and discharge	28,404	198,060	94,252
Age ≥ 18	28,404	190,509	94,252
Length of stay ≥ 6 h	27,848	177,162	93,216
First stay per admission	26,127	155,462	84,701
MAP and lactate within 24 h	3,868	54,524	4,228
48-hour trajectories used	**1,821**	**42,814**	**3,039**

Splits and cohort statistics. Stays are split into 80% training, 10% validation and 10% test within each hospital, with a fixed seed, so that every split contains all three hospitals and hours of one stay never appear in different splits. All models use the same split. The interventional metrics are computed on 450 test stays, 150 drawn at random from the test split of each hospital, so that every model is evaluated on the same patients. Table 5 summarises the cohorts.

Table 5: Cohort statistics over the first 48 hours. Percentages of stays unless stated otherwise.

	NWICU	eICU	MIMIC-IV
Training stays	1,457	34,252	2,433
Validation stays	182	4,281	303
Test stays	182	4,281	303
Vasopressor given	43.0	22.6	70.0
Hours on a vasopressor (% of hours)	13.5	8.5	43.2
Fluid bolus given	39.1	29.2	24.7
Mean MAP (mmHg)	82.2	79.9	75.9
Hours with MAP measured (%)	76.9	82.6	51.9
Mean lactate (mmol/L)	2.66	2.63	3.35
All measured cells (%)	15.2	13.8	11.2

B.2 Harmonisation

Shared schema. The three databases use different tables, identifiers, units and item codes. We map all of them to a shared hourly schema of 46 concepts (Table 6). The mapping step uses a language model on schema information only, such as column names, units and item labels, and every proposed mapping is checked against the source dictionary before use. Each concept is stored at every hour as its standardised value, which is zero when nothing was measured, an indicator of whether it was measured, and the logarithm of one plus the hours since its last measurement. The model therefore sees which values are missing and for how long.

Table 6: The 46 shared concepts.

Group	Count	Concepts
Vital signs	8	heart rate, SBP, MAP, DBP, respiratory rate, SpO_2, temperature, GCS
Blood gas	7	pH, $PaCO_2$, PaO_2, FiO_2, PaO_2/FiO_2, base excess, bicarbonate
Chemistry	11	sodium, potassium, chloride, calcium, magnesium, phosphate, glucose, BUN, creatinine, albumin, lactate
Liver and inflammation	2	total bilirubin, CRP
Blood and coagulation	6	haemoglobin, haematocrit, platelets, WBC, INR, PTT
Output	1	urine output
Organ failure	7	SOFA total and its six organ components
Admission	4	age, sex, admission weight, ICU admission type

Quality checks. Harmonising three schema families exposed several defects that produce plausible numbers rather than errors. Table 7 lists them together with the automatic check that now catches each one.

Table 7: Defects found during harmonisation and the checks added to catch them.

Defect	Consequence	Check
Lab events keyed by hospital admission, ICU events by ICU stay	A direct join dropped every lab value (16.7M rows in NWICU)	Row counts reported at every join
Time stored as timestamps, minute offsets or hours	A wrong unit shrank a median stay of 94.9 h to 0.026 h	Unit check on stay length
Temperature in Fahrenheit in one cohort	Temperature alone identified the hospital	Plausible range per concept
Placeholder 9,999,999 for "not reported"	Dominated the pooled standard deviation	Plausible range per concept
Drugs identified by integer item codes	Item 220949 (dextrose 5%) proposed as norepinephrine	Every code checked against the source dictionary

B.3 Treatment

At each hour the treatment is a binary vector with four channels: vasopressor, fluid bolus, mechanical ventilation and renal replacement therapy. A channel is 1 if an administration of that kind is recorded in the hour. The vasopressor channel covers norepinephrine, epinephrine, dopamine, phenylephrine and vasopressin; for reference, doses are converted to norepinephrine equivalents with factors of 1 for norepinephrine and epinephrine, 1/100 for dopamine, 1/10 for phenylephrine and 2.5 per unit/min for vasopressin. Mechanical ventilation is recorded in MIMIC-IV only, and renal replacement therapy is not recorded in these extracts. All four channels are model inputs. Our analysis concerns the vasopressor channel and its effect on MAP.

C Implementation

This section describes the prediction model, the construction of the contrast targets, the training settings and the ICU baselines.

C.1 Model

Inputs. At each hour, every concept is embedded from its value, measurement indicator and hours since last measurement, plus a learned identity vector shared by all hospitals. The embeddings are pooled over concepts with attention that down-weights, but does not remove, unmeasured concepts, which gives one vector f_t per hour.

Treatment input. The treatment vector a_t is added to the input and also applied through FiLM conditioning before and after the transformer:

$$\begin{aligned} \mathrm{FiLM}(h; a) &= \big(1 + \gamma(a)\big) \odot h + \beta(a), \\ z_t &= \mathrm{FiLM}\Big(\mathrm{Transformer}\big(\mathrm{FiLM}(W f_t + V a_t;\, a_t)\big);\, a_t\Big), \end{aligned} \tag{14}$$

where γ and β are linear maps initialised at zero, so that training starts from the additive model. The transformer is causal, so z_t depends only on hours up to t.

Outputs. From z_t, linear heads predict for every concept c the mean $\mu_{t+1,c}$ and variance $\sigma^2_{t+1,c}$ of its next value and the probability $\pi_{t+1,c}$ that it will be measured. The factual objective is the Gaussian negative log-likelihood on measured values plus the cross-entropy of the measurement indicator,

$$\mathcal{L}_{\mathrm{fact}} = \sum_{i,t,c} \Big[\tfrac{1}{2}\, m_{t+1,c} \Big(\log \sigma^2_{t+1,c} + \tfrac{(\tilde{x}_{t+1,c} - \mu_{t+1,c})^2}{\sigma^2_{t+1,c}} \Big) + \mathrm{BCE}\big(\pi_{t+1,c},\, m_{t+1,c}\big) \Big]. \tag{15}$$

Rollout. A rollout starts from 12 observed hours and feeds its own predictions back hour by hour, weighting each predicted value by its predicted measurement probability so that the rollout stays differentiable. To compute the contrast, we set the vasopressor channel to 1 at every hour and the other channels to 0, repeat with all channels at 0, and average the difference in predicted MAP over the next 12 hours. Table 8 summarises the architecture.

Table 8: Model architecture.

Component	Setting
Concept embedding	3-input MLP to 128 dimensions, plus concept identity vector
Transformer	6 causal layers, width 384, 6 heads, feed-forward 1536, dropout 0.1
Treatment input	additive, plus FiLM before and after the transformer
Output heads	mean, log-variance and measurement probability per concept
Parameters	10.8M

C.2 Matching Similar Patients

Propensity model. A one-layer GRU with 256 hidden units reads the values, measurement indicators and previous treatments, and predicts the four treatment channels at every hour.

It is trained on the training split for 30 epochs with Adam (learning rate 10^{-3}, batch 256) and then frozen.

Targets. Matching compares stays over the scored window, hours 13 to 24. A stay is treated if a vasopressor is given at any hour of the window, its outcome Y is the mean measured MAP in the window, and its propensity e is the mean predicted probability of a vasopressor over the window. At each hour the propensity model reads the record up to that hour, so the score summarises how strongly the patient's course over the window called for a vasopressor; it serves only to find patients with a similar course, and its dependence on the window is absorbed into the matching error of Assumption 2. This window-averaged score is a design choice rather than a standard pre-treatment propensity score; clinical practice and earlier exposure can influence it. Algorithm 1 builds the targets.

Treatment course. A target measures the course of vasopressor treatment that was actually given in the window, and this course is mostly sustained: 59% of treated training stays were already receiving a vasopressor in the hour before the window, and treated stays receive it for a median of 7 of the 12 hours. The model contrast instead switches the vasopressor on for all 12 hours. The contrast loss uses the targets only after removing their batch mean, so their level, which the length of the course affects most, never enters training; the direction of the effect is set by balancing. A shorter course shrinks a patient's target towards the shared offset, and Appendix A.3 shows that this cannot explain the dependence of the effect on pre-treatment pressure.

Algorithm 1 Building the contrast targets.

Require: training stays with treatment A_i, outcome Y_i and propensity e_i; k, δ

for each stay i with a measured outcome **do**
  $\mathcal{M}(i) \leftarrow$ the k stays with $A_j = 1 - A_i$ closest to e_i, keeping only $|e_j - e_i| \le \delta$
  if $|\mathcal{M}(i)| < k/5$ **then**
    skip stay i ▷ no comparable patients
  end if
  $\tilde{\tau}_i \leftarrow (2A_i - 1)\big(Y_i - \frac{1}{|\mathcal{M}(i)|}\sum_{j\in\mathcal{M}(i)} Y_j\big)$
end for

With $k = 10$ and $\delta = 0.05$, 36,833 of the 38,142 training stays have a measured outcome (7,885 treated and 28,948 untreated), and 31,690 of them have enough similar patients. The targets have mean -5.28 mmHg and standard deviation 13.5 mmHg. During training, each step draws 64 targets at random, with Huber threshold $\kappa = 1$ and scale s equal to the standard deviation of the centred targets, bounded below by 1.

C.3 Training and Hyperparameters

All ICU models are trained with AdamW, a linear warm-up over the first 5% of steps followed by cosine decay, gradient clipping at 1.0 and mixed precision, for at most 60 epochs with early stopping on the validation factual loss. The balancing classifier is a two-layer MLP with 128 hidden units, connected through gradient reversal. Its weight λ increases linearly over the first 10 epochs, and no checkpoint is selected during this period; both rules apply to every model that uses balancing, including CRN. Table 9 lists the hyperparameters of CauTwin.

Table 9: Hyperparameters of CauTwin on ICU data.

Group	Setting	Value
Optimisation	learning rate / weight decay	3×10^{-4} / 0.01
	batch size (without contrast supervision)	32 (128)
	maximum epochs / patience	60 / 8
Objective	weights λ / η / μ	3 / 1 / 1
	warm-up of λ	10 epochs
Contrast supervision	similar patients k / maximum gap δ	10 / 0.05
	observed hours / window W	12 / 12
	targets per step / Huber threshold κ	64 / 1

Each configuration is trained with 3 or 5 seeds. On one GPU, a model without contrast supervision trains in about 6 minutes and a model with it in about 2.3 hours, since every step runs two differentiable rollouts of 12 hours.

C.4 Baselines

All ICU baselines use the same data, split, outputs and training rules as CauTwin, and have between 9.84M and 11.15M parameters, within 9% of CauTwin's 10.78M.

- **GRU** (9.84M). A recurrent network trained on the factual objective.
- **RSSM** (10.71M). A recurrent state-space model with deterministic and stochastic states (Hafner et al., 2020), trained on the factual likelihood.
- **CRN** (9.84M) (Bica et al., 2020b). A recurrent encoder with adversarial balancing. We sweep $\lambda \in \{0.3, 1, 3, 10, 30\}$ and report $\lambda = 3$, the value used by CauTwin, and $\lambda = 0.3$, its best setting.
- **IPW** (10.78M). Our backbone with each hour weighted by its stabilised inverse propensity, as in RMSN (Lim et al., 2018), using the propensity model of Appendix C.2 with weights clipped at 10.
- **G-computation** (11.15M). Our backbone with additional heads that regress the model's own predictions under the intervened treatment step by step, as in IGC-Net (Hess et al., 2026a).

C.5 Evaluation Metrics

This subsection defines every metric reported in the paper. All ICU metrics are computed on the same 450 test patients. Each configuration is trained with 3 or 5 seeds, and we report the mean and a 95% confidence interval over seeds, $\bar{x} \pm t_{0.975,\,n-1}\, s/\sqrt{n}$, where n is the number of seeds and s the standard deviation across seeds.

Predicted effect. For a test patient i, we roll the model forward from the first 12 observed hours twice, once with the vasopressor channel set to 1 at every hour and once with all treatment channels set to 0, and record the predicted MAP $\hat{X}_{i,t}^{(1)}$ and $\hat{X}_{i,t}^{(0)}$ over the next $W = 12$ hours. The predicted effect of patient i is

$$\hat{\tau}_i = \frac{1}{W} \sum_{t=13}^{24} \left(\hat{X}_{i,t}^{(1)} - \hat{X}_{i,t}^{(0)} \right) \tag{16}$$

in mmHg, and the column "Effect" in Tables 1 and 3 is its average over the $n = 450$ test patients.

Correct direction. A vasopressor raises MAP, so the predicted effect should not be negative. Correct direction is the share of test patients for which it is not,

$$\mathrm{CD} = \frac{100}{n} \sum_{i=1}^{n} \mathbf{1}\left[\hat{\tau}_i \geq 0\right]. \tag{17}$$

It checks the level of the contrast. A model that confuses treatment with severity predicts negative effects and scores below 50%.

Patient dependence. Patients with lower pressure should gain more from a vasopressor. Let m_i be the mean of the measured MAP values over the first 12 hours of patient i. Patient dependence is the Spearman rank correlation between the predicted effects and these pre-treatment values,

$$\rho = \mathrm{corr}\big(\mathrm{rank}(\hat{\tau}_i),\, \mathrm{rank}(m_i)\big), \tag{18}$$

computed over patients with at least one measured MAP value. It checks the shape of the contrast: a negative value means that the effect decreases with pre-treatment pressure, and a value near zero means that the effect ignores the patient.

Other outcomes and hospitals. The same two metrics are computed for SBP and DBP, which a vasopressor should also raise, and separately for the 150 test patients of each hospital. For heart rate we report the mean predicted effect in beats per minute.

Rollout error. To measure predictive accuracy, we roll the model forward from the first 12 observed hours under the recorded treatments and compare the predicted MAP with the measured values,

$$\mathrm{RMSE} = \Big(\frac{1}{|\mathcal{O}|}\sum_{(i,t)\in\mathcal{O}}\big(\hat{X}_{i,t} - X_{i,t}\big)^2\Big)^{1/2}, \tag{19}$$

where $\mathcal{O}$ is the set of patient hours with a measured MAP. Figure 5 uses a 24-hour rollout on the full test split, and Figure 7b uses a 12-hour rollout on the full test split, split by whether the patient received a vasopressor.

Physiological consistency. The violation rate is the share of predicted hours in which the ordering SBP $\geq$ MAP $\geq$ DBP fails. It is computed under the recorded treatments and averaged over seven treatment sequences that were not given (Appendix D.1).

Tumour benchmark. The simulator provides the outcome under any treatment sequence. Counterfactual RMSE compares the predicted tumour volume with the simulated one, averaged over 50 random treatment sequences per test patient, for each horizon $h = 1, \ldots, 6$ and on average. For the treatment effect at horizon 3, with true effect τ_i and prediction $\hat{\tau}_i$,

$$\mathrm{PEHE} = \Big(\frac{1}{n}\sum_i(\hat{\tau}_i - \tau_i)^2\Big)^{1/2}, \qquad \text{ATE error} = \Big|\frac{1}{n}\sum_i \hat{\tau}_i - \frac{1}{n}\sum_i \tau_i\Big|, \tag{20}$$

and we also report the Pearson correlation between $\hat{\tau}_i$ and τ_i and the share of patients whose predicted effect has the correct sign.

D More ICU Results

This section reports per-seed results and further analyses on the same 450 test patients as the main text. Rollouts are sampled, so repeated evaluations of the same model differ by a few percentage points.

D.1 Ablation

Per-seed results. Table 10 lists every seed behind Table 3. CauTwin is the most stable configuration across seeds, with a correct direction between 79.3% and 80.2%.

Table 10: Correct direction (%) and patient dependence ρ for every seed.

Variant	Seed 0	Seed 1	Seed 2	Seed 3	Seed 4
Factual objective only	23.3 / −0.32	30.2 / −0.32	28.7 / −0.17	39.1 / −0.27	32.2 / −0.29
Balancing only	89.8 / +0.00	92.7 / +0.07	76.7 / +0.10		
Contrast supervision only	38.0 / −0.59	39.3 / −0.59	37.3 / −0.54		
CauTwin	80.2 / −0.51	79.6 / −0.44	79.3 / −0.44		
Stronger contrast supervision	58.0 / −0.56	74.7 / −0.56	79.1 / −0.49		
Balancing on part of the state	36.4 / −0.54	36.2 / −0.56	28.7 / −0.55		
Treatment as residual	40.2 / −0.41	41.1 / −0.47	38.2 / −0.53	40.2 / −0.50	37.6 / −0.54

Weight of contrast supervision. Figure 6 varies η with $\lambda = 3$. The patient dependence improves sharply from $\eta = 0$ to $\eta = 1$ and barely changes beyond, while the correct direction decreases slowly and becomes unstable at $\eta = 3$. We therefore use $\eta = 1$.

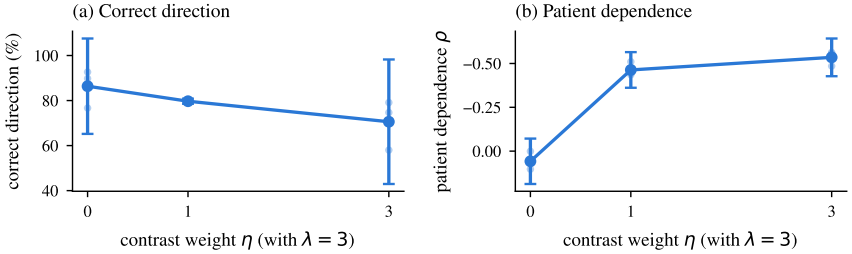


Figure 6: **Weight of contrast supervision.** Correct direction (a) and patient dependence (b) as η grows, with balancing fixed at $\lambda = 3$. Small dots are seeds; error bars are 95% confidence intervals.

Additive and FiLM inputs. With balancing only, an additive treatment input and FiLM conditioning give the same correct direction ($87.9 \pm 7.8\%$ and $87.1 \pm 20.1\%$, three seeds each), and neither makes the effect depend on the patient ($\rho = +0.08$ and $+0.00$). The extra flexibility of FiLM is used only when contrast supervision asks for it, as Corollary 1 predicts.

Physiological term. Table 11 reports how often predicted rollouts violate SBP $\geq$ MAP $\geq$ DBP, both under the observed treatments and averaged over seven treatment sequences that were not given. The physiological term keeps violations close to zero in both cases, and removing it does not change the correct direction (82.9% at $\lambda = 3$).

Table 11: Violations of SBP $\geq$ MAP $\geq$ DBP in predicted rollouts.

Model	Observed treatments (%)	Treatments not given (%)
Balancing, without physiological term	2.23	1.27
Balancing, with physiological term	0.00	0.02
CauTwin (three seeds)	0.00 to 0.05	0.00 to 0.01

D.2 Balancing

Input to the classifier. A linear probe predicts the current treatment from the internal state equally well with and without balancing (AUC 0.975 and 0.976 on the state of the previous hour), so balancing does not act by removing treatment information from the state. To locate its effect, Table 12 changes what the classifier reads. Reading the previous hour, as in CRN (Bica et al., 2020b), works as well as the standard classifier and again combines with contrast supervision. Reading a state computed with the treatment inputs set to zero removes the effect of balancing. Balancing therefore acts on how the model uses its treatment inputs, preventing them from standing in for severity.

Table 12: Balancing with different inputs to the classifier ($\lambda = 3$).

Classifier reads	Correct direction (%)	Patient dependence ρ
Current hour (standard)	85.7 ± 6.7	-0.02 ± 0.14
Previous hour	88.7 ± 6.6	-0.09 ± 0.11
Treatment inputs set to zero	22.2 ± 7.5	-0.12 ± 0.04
Previous hour, with contrast supervision	82.9 ± 1.0	-0.48 ± 0.13

Separating the two terms. We tested two designs that keep severity away from the treatment pathway without balancing the whole model. In the first, the internal state is split

into two parts, the output is the sum of a term that reads one part and a term that reads the other part together with the treatment, and only the second part is balanced. In the second, the treatment enters as its difference from the predicted propensity, as in the R-learner (Nie & Wager, 2021), with no classifier. Both learn the shape but not the direction: the split design reaches 27.6% correct direction alone and 33.8% with contrast supervision, and the residual design 39.5% with contrast supervision (Tables 3 and 10). Applying both terms to the whole model, as CauTwin does, is what makes the two properties compatible.

D.3 Generalisation

Other outcomes and hospitals. Figure 7a evaluates CauTwin on outcomes it was not tuned for and on each hospital separately. The correct direction holds for systolic (86.4%) and diastolic pressure (78.4%) and in every hospital (73.1% in NWICU, 78.7% in eICU and 89.8% in MIMIC-IV). Every seed also predicts a fall in heart rate after a vasopressor (-2.2, -5.1 and -0.5 beats per minute), the baroreflex response to a rise in pressure.

Cost in prediction accuracy. Figure 7b splits the MAP rollout error by whether the patient received a vasopressor, on the full test split (1,221 treated and 3,545 untreated patients, five seeds). With balancing, the error rises from 13.61 to 15.87 mmHg on treated patients and changes only from 14.39 to 14.51 mmHg on untreated ones, which is where the shortcut "treated, so severe, so low pressure" is removed.

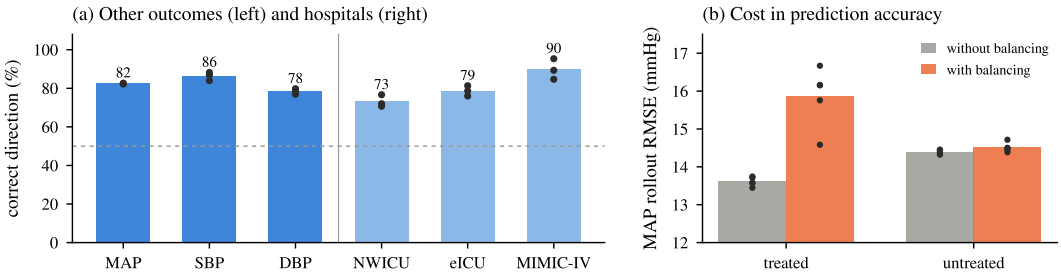


Figure 7: **Generalisation and cost.** (a) Correct direction of the vasopressor effect for CauTwin on other blood pressures and in each hospital. (b) MAP rollout RMSE on the full test split for models with and without balancing, split by whether the patient received a vasopressor. Dots are seeds.

D.4 Model Selection

Table 13 trains CauTwin with balancing only and CRN at several balancing weights. For our model, the validation loss rises with λ while the correct direction also rises, so selecting by validation loss would pick $\lambda = 0$, the worst setting on the interventional question; we use $\lambda = 3$, where the direction is close to its maximum. For CRN, both criteria choose $\lambda = 0.3$, which still reaches only 55.8%. Across all 119 ICU models, the correlation between MAP rollout error and correct direction is +0.48 (Figure 5a).

Table 13: Validation loss and correct direction across balancing weights (seed 0).

	λ	0	0.3	1	3	10	30
Ours	Validation loss (↓)	−0.005	−0.003	0.006	0.016	0.037	0.052
	Correct direction (%) (↑)	23.3	52.9	81.1	92.4	93.1	94.0
CRN	Validation loss (↓)		−0.031	−0.024	0.010	0.094	0.288
	Correct direction (%) (↑)		55.8	40.9	32.2	37.8	40.4

E Tumour Benchmark

This section specifies the semi-synthetic benchmark, the model configuration used on it, and the full results behind Table 2.

E.1 Setup

Simulator. We use the tumour growth simulator of Bica et al. (2020b), based on Gompertz growth (Benzekry et al., 2014). Each patient has a tumour volume V_t, a chemotherapy concentration C_t and an accumulated radiation dose R_t, and receives two binary treatments a_t^{c} and a_t^{r} at each step:

$$V_{t+1} = V_t + \rho_p V_t \log\big(K/V_t\big) - \beta_c C_t V_t - \big(\alpha_r R_t + \beta_r R_t^2\big) V_t + \epsilon_t, \tag{21}$$

$$C_{t+1} = C_t\, e^{-\lambda_c} + a_t^{\mathrm{c}}, \qquad R_{t+1} = R_t\, e^{-\lambda_r} + a_t^{\mathrm{r}}, \tag{22}$$

with multiplicative noise ϵ_t. Patients belong to one of four unobserved types that differ in growth rate ρ_p, so the effect of chemotherapy differs between patients.

Confounding. Treatment is more likely for larger tumours and less likely when the drug level is already high,

$$\mathbb{P}\left(a_t^{\mathrm{c}} = 1 \mid V_t, C_t\right) = \sigma\Big(\gamma\, \frac{V_t - V_{\mathrm{ref}}}{V_{\mathrm{ref}}} - C_t\Big), \tag{23}$$

and radiotherapy is assigned in the same way with R_t in place of C_t. The parameter γ sets the strength of confounding. Tables 14 and 15 give the simulator parameters and the resulting strength of confounding.

Table 14: Simulator parameters.

Symbol	Meaning	Value
T	trajectory length	60
N	patients per setting	5,000
ρ_p	growth rate per type	$\mathcal{U}(0.04, 0.06)$
K	carrying capacity	30
V_0	initial volume	$\mathcal{U}(0.5, 5.0)$
β_c	chemotherapy kill rate	0.005
α_r, β_r	radiotherapy coefficients	0.005, 0.0005
λ_c, λ_r	elimination rates	0.139
σ_V	relative noise	0.01

Table 15: Strength of confounding.

γ	Treated	$\mathrm{corr}(V_t, a_t)$	Oracle AUC
1	0.42	0.13	0.68
2	0.54	0.19	0.70
3	0.62	0.23	0.71
4	0.68	0.26	0.73

Evaluation. For each γ and seed we generate 5,000 patients and split them 70/15/15. Each test patient is observed for 10 steps and then rolled out under 50 random treatment sequences of length up to 6. The ground truth is obtained by re-simulating from the same state under the same sequence, averaged over 50 noise draws, and we report the RMSE of tumour volume.

Baselines. We compare with eleven methods at the settings of their original papers (Table 16). All use AdamW with learning rate 3×10^{-4}, weight decay 0.01, batch size 128 and early stopping with patience 10 within at most 80 epochs, except that ACTIN uses the learning rates of its original paper.

Table 16: Tumour benchmark models.

Method	Architecture	Parameters
GRU	2-layer GRU	0.22M
CRN	2-layer GRU, balancing	0.22M
RMSN	2-layer GRU, inverse propensity weights	0.29M
G-Net	2-layer GRU, G-computation	0.29M
CT	3-layer transformer, balancing	1.31M
IGC-Net	3-layer transformer, iterated regression	1.52M
WO-Learner	3-layer transformer, overlap weights	0.63M
ACTIN	temporal convolutions, balancing	0.14M
TE-CDE	neural CDE	0.68M
Causal CPC (El Bouchattaoui et al., 2024)	contrastive predictive coding, balancing	0.15M
Mamba-CDSP (Gu & Dao, 2024)	state-space model, balancing	0.12M
CauTwin	4-layer transformer (Appendix E.2)	1.02M

E.2 Model

The simulator has two treatments, a low-dimensional state and short trajectories, so we use a lighter configuration of CauTwin adapted to it. The encoder reads past outcomes only, and for each horizon $k = 1, \dots, 6$ a separate head predicts the untreated outcome and the effect of each treatment,

$$\hat{V}_{t+k} = f_k(z_t) + \sum_{j \in \{\mathrm{c},\mathrm{r}\}} a_{t,j}\, \mathrm{effect}_{k,j}(z_t). \tag{24}$$

Training has two stages. In the first (56 epochs), all parameters fit the observed outcomes; on treated steps, an effect-boost loss stops the gradient through the base head, so that the effect head must explain what the base head does not. In the second (40 epochs), the encoder and base heads are frozen, a propensity head $\hat{e}$ is fitted and clipped to $[0.1, 0.9]$, and the effect heads minimise $\big(y - f(z) - (a - \hat{e}(z))\,\mathrm{effect}(z)\big)^2$, as in the R-learner (Nie & Wager, 2021). In the simulator, treatment depends only on the recorded tumour volume and the past treatments, so there is no unrecorded severity for contrast supervision to correct. The benchmark therefore evaluates the prediction head, rollout and effect training of CauTwin, while the contrast supervision of Section 4.2 is evaluated on ICU data, where the confounding it addresses arises.

E.3 Results

Full results. Table 17 extends Table 2 with all twelve methods and standard deviations. CauTwin has the lowest average error.

Table 17: Counterfactual RMSE on the tumour benchmark, mean ± std over three seeds.

Method	γ=1	γ=2	γ=3	γ=4	Avg
CauTwin	0.434 ± 0.037	0.364 ± 0.068	0.330 ± 0.034	0.329 ± 0.039	**0.364**
IGC-Net	0.455 ± 0.022	0.373 ± 0.017	0.353 ± 0.030	0.334 ± 0.037	0.379
CT	0.465 ± 0.017	0.392 ± 0.044	0.366 ± 0.028	0.343 ± 0.032	0.392
G-Net	0.445 ± 0.025	0.395 ± 0.025	0.368 ± 0.023	0.366 ± 0.027	0.394
RMSN	0.433 ± 0.019	0.398 ± 0.028	0.392 ± 0.024	0.381 ± 0.033	0.401
CRN	0.486 ± 0.053	0.418 ± 0.061	0.395 ± 0.066	0.363 ± 0.055	0.416
GRU	0.488 ± 0.008	0.412 ± 0.033	0.404 ± 0.044	0.398 ± 0.058	0.426
ACTIN	0.575 ± 0.072	0.461 ± 0.075	0.423 ± 0.063	0.430 ± 0.065	0.472
WO-Learner	0.555 ± 0.019	0.466 ± 0.034	0.447 ± 0.038	0.443 ± 0.053	0.478
Causal CPC	0.647 ± 0.059	0.520 ± 0.054	0.437 ± 0.017	0.418 ± 0.048	0.505
Mamba-CDSP	0.652 ± 0.114	0.548 ± 0.036	0.491 ± 0.027	0.498 ± 0.086	0.547
TE-CDE	1.213 ± 0.165	0.953 ± 0.059	0.829 ± 0.033	0.731 ± 0.058	0.932

Error by horizon. Figure 8 shows that the methods are close at one step and separate as the rollout grows. CauTwin has the lowest error from $h = 4$ onwards, rising from 0.291 at $h = 1$ to 0.491 at $h = 6$. Mamba-CDSP has the lowest one-step error (0.267) but reaches 0.873 at $h = 6$.

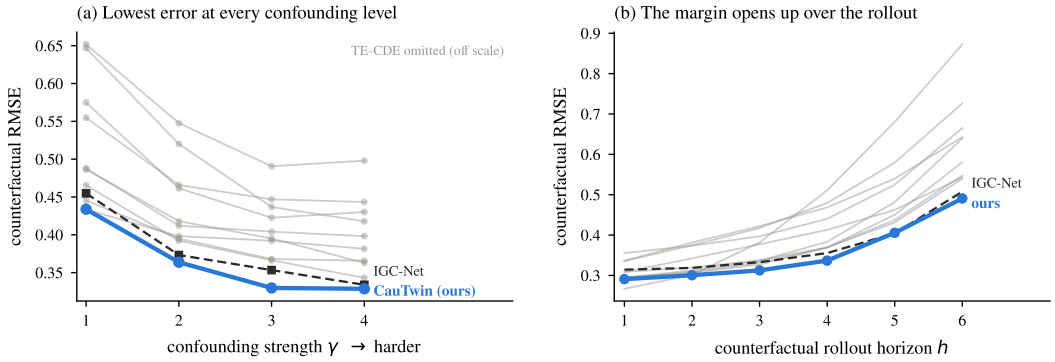


Figure 8: **Counterfactual error on the tumour benchmark.** (a) By confounding strength. (b) By rollout horizon, averaged over γ. Grey lines are the other baselines.

Effect estimates. Table 18 reports point estimates of the effect at horizon 3. The Causal Transformer is strongest on these single-horizon metrics, and CauTwin ranks third on correlation, consistent with its configuration being tuned for multi-step rollout. This configuration has no contrast supervision, the component that shapes individual effects on ICU data, so these numbers describe its effect heads rather than the contribution of this paper.

Table 18: Effect estimates at $h = 3$, averaged over γ and seeds.

Method	PEHE (↓)	ATE error (↓)	Correlation (↑)	Correct sign (↑)
CT	**0.340**	**0.277**	0.698	**0.776**
IGC-Net	0.381	0.289	0.022	0.773
CRN	0.383	0.302	0.322	0.771
G-Net	0.451	0.402	0.009	0.596
CauTwin	0.453	0.392	0.624	0.615
GRU	0.498	0.438	0.240	0.538
RMSN	0.541	0.494	−0.116	0.404
WO-Learner	0.630	0.435	0.152	0.512
Causal CPC	0.632	0.510	−0.088	0.429
ACTIN	0.648	0.616	0.225	0.343
Mamba-CDSP	0.796	0.688	−0.066	0.232
TE-CDE	1.251	1.182	**0.705**	0.002

Ablation. Table 19 removes one part of the configuration at a time. Each removal raises the average error by more than the margin to the second-best method, and Figure 9 shows that the cost grows with the horizon. One run of CauTwin, including evaluation, takes 79 seconds on one GPU.

Table 19: Ablation on the tumour benchmark: RMSE, mean ± std over three seeds.

Variant	γ=1	γ=2	γ=3	γ=4	Avg	Change
Full configuration	0.434 ± 0.037	0.364 ± 0.068	0.330 ± 0.034	0.329 ± 0.039	**0.364**	
Without effect boost	0.632 ± 0.042	0.491 ± 0.031	0.478 ± 0.054	0.471 ± 0.061	0.518	+0.154
Without R-learner stage	0.581 ± 0.030	0.501 ± 0.052	0.480 ± 0.050	0.472 ± 0.061	0.508	+0.144
Encoder reads treatments	0.607 ± 0.021	0.495 ± 0.020	0.453 ± 0.012	0.435 ± 0.020	0.497	+0.133

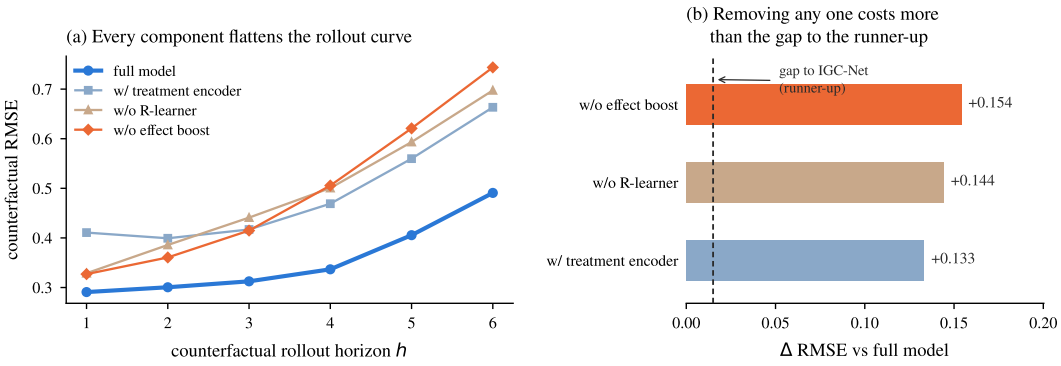


Figure 9: **Ablation on the tumour benchmark.** (a) Error by horizon for each variant. (b) Increase in average error when each part is removed.

F Scope and Future Work

This section summarises the scope of our results and the directions they open.

Identification assumptions. Like other methods for observational data, CauTwin relies on sequential ignorability (Assumption 1). Contrast supervision adds only a mild requirement on top of it: similar patients may differ from each patient by any offset, as long as the offset is shared (Assumption 2), and Proposition 2 shows that the error in the shape grows only with how much this offset varies. Appendix A.3 checks this across hospitals, severity and the course of treatment; a formal sensitivity bound for offsets that vary with unrecorded severity is a natural extension. Related approaches address failures of unconfoundedness by partially identifying effects (Schweisthal et al., 2024; 2025); our experiments do not provide such bounds.

Evaluation on real data. Counterfactual outcomes are never observed in ICU data. We therefore evaluate two properties that pharmacology fixes in advance, the direction of the effect and its dependence on pre-treatment pressure, and complement them with a semi-synthetic benchmark where the ground truth is known. Validating effect sizes against randomised or prospective data is the next step. Uncertainty quantification for dynamic treatment regimes is another relevant extension (Deng et al., 2025).

Treatment definition. We study vasopressor treatment over a 12-hour window as a binary decision, one of the most common haemodynamic interventions in the ICU. Matching on the length of the course, or on a propensity computed before the window, would align the targets more closely with the sustained course that we evaluate. Extending the contrast to doses, timing and combinations of treatments would require rechecking the matching assumptions. Recent work studies time-varying clinical effects (Chien et al., 2024) and intervention-sequence planning (Wang et al., 2025).

Strength of patient dependence. CauTwin is the only model whose effect depends on the patient in the pharmacologically expected way, and part of its effect is still shared across patients. Strengthening this dependence, for example by conditioning the treatment pathway more directly on physiological state, is a promising direction.

G Related Work

This section extends the related work of the main text.

G-methods. Treatment effects under time-varying confounding were first formalised by the G-methods of Robins (1986) and Robins et al. (2000): G-computation, marginal structural models fitted with inverse probability weighting, and structural nested models (Hernán & Robins, 2020). These methods target effects averaged over a population.

Neural methods for treatment effects over time. Neural methods carry the same ideas to high-dimensional patient histories. RMSN (Lim et al., 2018) learns propensity networks and reweights observations by their inverse probability. CRN (Bica et al., 2020b) and the Causal Transformer (Melnychuk et al., 2022) learn a representation of the history from which the treatment cannot be predicted, using adversarial training. ACTIN (Wang et al., 2024) follows the same balancing idea, and TE-CDE (Seedat et al., 2022) combines it with neural controlled differential equations (Kidger et al., 2020) to handle irregular sampling. G-Net (Li et al., 2021) performs G-computation by simulating future covariates, and IGC-Net (Hess et al., 2026a) replaces this simulation with iterated regression. Representation balancing for static counterfactual prediction (Johansson et al., 2016; Shalit et al., 2017) can introduce bias when the representation discards confounders (Melnychuk et al., 2024; Huang et al., 2024). These architectures generally produce treatment effects from their estimated potential outcomes. Causal CPC (El Bouchattaoui et al., 2024) uses contrastive predictive coding (van den Oord et al., 2018) for longer-horizon representations, rather than directly supervising patient-level interventional contrasts. Deep LTMLE (Shirakawa et al., 2024) targets the longitudinal counterfactual mean through a correction step.

Meta-learners. Meta-learners estimate the effect with a second model fitted to pseudo-outcomes whose expectation equals the effect. The DR-learner (Kennedy, 2023) and the R-learner (Nie & Wager, 2021) are the standard constructions for static data, and recent work extends them to time-varying settings with model-agnostic pseudo-outcomes (Frauen et al., 2025), overlap weighting (Hess et al., 2026b) and orthogonal estimation of Q-functions (Javurek et al., 2026). Inverse-variance weights can stabilize static pseudo-outcome regressions (Fisher, 2024). Counterfactual alignment at the sample level (Zhang et al., 2025) tackles a related heterogeneity problem for static treatments. Our target is close in spirit to a pseudo-outcome, but it is built by propensity matching (Rosenbaum & Rubin, 1983), only its differences between patients are used, and it supervises the effect of the same model that predicts the trajectory. Table 20 summarises these differences.

Table 20: How existing methods for treatment effects over time relate to CauTwin.

Method	Handles confounding by	Effect supervised per patient	One model for trajectory and effect
GRU, RSSM	nothing	✗	✓
CRN, CT, ACTIN	balanced representation	✗	✓
TE-CDE	balancing in continuous time	✗	✓
RMSN	inverse propensity weighting	✗	✓
G-Net, IGC-Net	G-computation	✗	✓
DR-, R-learner	pseudo-outcomes	✓	✗
WO-, DRQ-learner	orthogonal pseudo-outcomes	✓	✗
CauTwin (ours)	balancing and matched patients	✓	✓

Learned models for clinical decisions. Model-based and offline reinforcement learning has been proposed to recommend ICU treatments, with transition models learned from records (Komorowski et al., 2018; Wang et al., 2018; Tang & Wiens, 2022). Recent intervention planning considers the compounding error from missing counterfactuals (Wang et al., 2025). The counterfactual predictions of such models cannot be checked against data, and small modelling choices can change the recommended policy (Gottesman et al., 2019; Jeter et al., 2019). Model selection under missing counterfactual outcomes has been studied separately (Curth & van der Schaar, 2023; Gao, 2025). Section 5.4 adds that the usual criterion for choosing these models, accuracy on the observed future, favours models whose interventional answers point in the wrong direction.