http://www.diva-portal.org

This is the published version of a paper presented at *International Conferenceon AI in Medicine and Healthcare*.

Citation for the original published paper:

Khan, A S., Schaffernicht, E., Stork, J A. (2026)
Policy-Focused Evaluation of Individualized Vasopressor Effects with Robustness to Irrelevant Covariates in MIMIC-III ICU Data
In: *Proceedings of the 2nd International Conferenceon AI in Medicine and Healthcare (AiMH' 2026)* Barcelona

N.B. When citing this work, cite the original published paper.

Permanent link to this version:
http://urn.kb.se/resolve?urn=urn:nbn:se:oru:diva-131458

2nd International Conference on AI in Medicine and Healthcare (AiMH' 2026)
24-26 June 2026, Palma de Mallorca, Spain

In-Person: Oral ☐ / Poster ☐ / The same ☐
Virtual: Zoom ☒ / Pre-recorded video ☐
Topic: *AI in Patient Care and Personalized Medicine*

Policy-Focused Evaluation of Individualized Vasopressor Effects with Robustness to Irrelevant Covariates in MIMIC-III ICU Data

A. S. Khan [1], E. Schaffernicht [2] and J. A. Stork [1]
[1] School of Science and Technology, Örebro University, Sweden
[2] Technical University of Applied Sciences Würzburg-Schweinfurt, Germany
E-mail: {ahmad-saeed.khan, johannesandreas.stork}@oru.se, erik.schaffernicht@thws.de

Summary: Estimating individualized treatment effects (ITEs) from observational intensive care unit (ICU) data is challenging because treatment assignment is strongly confounded by patient severity, overlap between treatment groups may be limited, and high-dimensional clinical covariates often contain irrelevant measurements that degrade treatment-effect models. We study individualized vasopressor effects using the MIMIC-III ICU database, where treatment is defined as vasopressor administration within the first 24 hours of ICU admission. The outcome $y_i$is a clinically motivated lactate-change endpoint defined as the difference between mean serum lactate measured in a late post-admission window (60–84h) and a baseline window around ICU admission (−6h to +6h), where negative values indicate improvement. Because counterfactual outcomes are unobserved, we adopt a policy-focused evaluation protocol combining model-implied average treatment effect (ATE) estimation with bootstrap uncertainty, plug-in policy-risk evaluation under a treat-if-benefit rule $\pi_f(x)$, heterogeneity diagnostics, subgroup analysis across baseline lactate strata, and robustness analysis under controlled irrelevant-covariate inflation. We compare DRI-ITE, which learns disentangled latent factors corresponding to treatment-predictive (Γ), confounding (Δ), adjustment (Υ), and irrelevant (Ω) variation, against standard meta-learners (T-, S-, and X-learners), a causal forest, and disentanglement-based baselines (DRCFR and RLOCFR). While model-implied ATE estimates are broadly similar across methods, policy evaluation reveals substantial differences in decision quality. DRI-ITE achieves the largest estimated improvement over a treat-all policy ($\Delta_{\text{all}} = -0.132 \pm 0.009$) while maintaining a balanced treatment rate, whereas several baselines collapse toward near-universal treatment with negligible policy gain. Under controlled irrelevant-feature inflation via appended Gaussian covariates ($|\ \Omega\ | \in \{5,10,15,20\}$), DRI-ITE consistently retains the strongest policy improvement and stable treatment behavior, while causal forest, S-learners, and weaker disentanglement-based baselines deteriorate toward more aggressive or less informative policies. These results demonstrate that evaluating ITE models through decision-focused criteria reveals failure modes that are not apparent from average-effect estimates alone, and that explicitly separating irrelevant covariate variation leads to more robust and clinically meaningful treatment policies in real-world observational settings.

Keywords: Individualized treatment effects, ICU, MIMIC-III, vasopressors, causal inference, policy learning.

1. Introduction

Vasopressors are a cornerstone of haemodynamic resuscitation in the intensive care unit (ICU): they increase arterial pressure and help maintain end-organ perfusion in critically ill patients. However, their benefit is highly heterogeneous. In some patients with severe vasodilatory shock, early vasopressor administration is essential for stabilizing circulation, whereas in others aggressive vasopressor use may be unnecessary or even harmful. Average treatment effect (ATE) analyses therefore provide limited guidance for clinical decision-making. Instead, clinicians need models that can estimate *individualized treatment effects* (ITEs) and identify which patients are most likely to benefit from treatment.

ITE estimation from observational data has been widely studied in causal machine learning, building on the potential outcomes framework[1],[2] representation learning approaches[3] and doubly robust estimation techniques[4],[5]. Applying these methods to ICU data introduces several challenges.

Confounding.

Treatment assignment in clinical practice depends strongly on patient covariates such as hemodynamic instability, laboratory measurements, and clinician judgment. Consequently, the propensity $P(T = 1 \mid X)$varies substantially across the covariate space, introducing selection bias relative to a randomized treatment setting.

Limited overlap.

Vasopressors are typically administered to severely ill patients and less frequently to stable ones. Regions of the covariate space with near-zero or near-one treatment propensity can lead to unstable estimation, especially for methods relying on propensity-based reweighting or adjustment.

High-dimensional nuisance covariates.

ICU datasets contain many physiological measurements and laboratory variables, some of which may be weakly related or irrelevant to treatment assignment and outcome prediction. Such nuisance

dimensions can inflate estimator variance, increase overfitting, and induce overly aggressive or degenerate treatment policies.

Existing empirical evaluations of ITE methods on clinical data often emphasize average-effect recovery or proxy prediction criteria that do not directly reflect the quality of individualized treatment decisions. In practice, however, the central question is whether a model can induce a useful treatment policy. Because counterfactual outcomes are not observed in observational data, the true value of a policy cannot be directly evaluated. Instead, we adopt a *policy-focused evaluation* strategy based on model-implied treatment rules, where policy value is estimated using predicted counterfactual outcomes. These quantities should therefore be interpreted as plug-in diagnostics that reflect the internal consistency and decision behavior of a model, rather than unbiased estimates of real-world clinical utility.

In this work, we study individualized vasopressor treatment effects using the real-world MIMIC-III intensive care unit database [6]. We construct a cohort of adult ICU patients, define treatment as vasopressor administration within the first 24 hours of ICU admission, and use a clinically motivated lactate-change endpoint comparing baseline lactate measurements around admission to mean lactate levels measured 60–84 hours later. Using this cohort, we evaluate heterogeneous treatment effect estimators under several complementary criteria.

Our analysis focuses on the empirical behaviour of DRI-ITE [7], a disentangled representation learning framework for ITE estimation that separates treatment-predictive, confounding, outcome-adjustment, and irrelevant covariate variation. We compare DRI-ITE not only against widely used meta-learners (T-, S-, and X-learners) and causal forests, but also against related disentanglement-based baselines.

While a wide range of ITE estimation methods have been proposed, we focus in particular on DRI-ITE because its design explicitly targets a key challenge in observational ICU data: the presence of high-dimensional covariates that include both clinically relevant and irrelevant information. In contrast to standard meta-learners and tree-based approaches, DRI-ITE separates covariate information into treatment-predictive, confounding, outcome-related, and nuisance components through a disentangled representation. This structure is particularly well suited to clinical settings, where many recorded variables may be weakly related or irrelevant to the treatment decision and outcome, and can therefore degrade model reliability and lead to unstable or overly aggressive treatment policies. The MIMIC-III ICU database provides a realistic and challenging setting for evaluating such models. Treatment assignment is strongly confounded by patient severity, overlap between treated and untreated patients is limited, and the covariate space includes a wide range of physiological measurements and laboratory variables with varying relevance. These characteristics make MIMIC-III an appropriate testbed for studying whether ITE models can produce stable and clinically meaningful treatment policies under realistic data conditions.

Our study makes four contributions.

- We provide a reproducible real-world evaluation of individualized vasopressor treatment effects using the MIMIC-III ICU database with a clinically motivated outcome based on lactate dynamics.
- We introduce a structured, policy-focused evaluation protocol for individualized treatment effect models in observational clinical data. The protocol integrates model-implied ATE estimation, plug-in policy evaluation, subgroup analysis, and robustness assessment under controlled irrelevant-covariate inflation, enabling a more comprehensive assessment of decision quality beyond standard predictive metrics.
- We demonstrate empirically that DRI-ITE produces more stable and informative treatment policies than standard meta-learners, causal forests, and related disentanglement-based baselines, particularly under high-dimensional nuisance variation, highlighting the importance of separating irrelevant covariates in real-world settings.
- We provide an interpretable analysis of the learned representations by examining how different groups of covariates contribute to treatment prediction and outcome estimation. This analysis shows that the model separates clinically meaningful patient characteristics from less relevant variation, supporting the interpretability of the learned structure.

Taken together, these results highlight the importance of explicitly accounting for irrelevant covariate structure when estimating individualized treatment effects from high-dimensional observational ICU data.

2. Related Work

Heterogeneous treatment effects from observational data.

Early approaches to heterogeneous treatment effect estimation relied on matching and propensity score adjustments within the potential outcomes framework [1],[8]. More recently, machine learning methods have been developed to estimate individualized treatment effects from high-dimensional data. Meta-learners such as the S-, T-, and X-learners [9] provide flexible plug-in estimators that can be implemented with arbitrary supervised learning models. Tree-based approaches such as causal trees and causal forests [10], [11] estimate local treatment effects using adaptive partitioning with asymptotic guarantees. Representation-learning methods aim to reduce confounding by learning balanced latent representations prior to outcome prediction, including TARNet and CFRNet [3], [12], variational models such as CEVAE [13] , and architectures such as DragonNet [14]. While these approaches achieve strong performance on semi-synthetic benchmarks, their behavior in real-world clinical settings with high-dimensional nuisance variation remains less well understood.

Disentangled representations for causal inference.

Recent work has explored disentangled representation learning as a strategy for improving treatment effect estimation. These approaches aim to separate covariate information related to treatment assignment from information predictive of outcomes in order to better control confounding [3], [15], [16]. By structuring latent representations into interpretable components, such models seek to isolate confounding variation and improve counterfactual prediction. However, empirical evaluations of disentangled causal representations have largely focused on benchmark datasets or proxy metrics, leaving open questions about their practical behavior under real-world data conditions. In this work, we study the empirical behavior of DRI-ITE [7], which explicitly factorizes treatment-predictive, confounding, outcome-related, and irrelevant covariate variation, and evaluate its performance under clinically motivated and robustness-focused criteria.

Causal inference with electronic health records.

Causal machine learning has increasingly been applied to electronic health records and ICU datasets. Publicly available resources such as the MIMIC-III critical care database [6] have enabled empirical studies of treatment strategies and counterfactual prediction in real-world clinical settings. However, evaluation remains challenging because counterfactual outcomes are not observed. Prior work therefore relies on proxy metrics such as PEHE [3], off-policy evaluation methods [17], or policy learning frameworks[18], [19]. Our work follows a similar policy-focused perspective but emphasizes practical evaluation of individualized treatment policies derived from observational ICU data, rather than relying solely on proxy prediction metrics.

Irrelevant covariates and robustness in causal estimation.

Including irrelevant or weakly related covariates can increase estimator variance and, in some cases, amplify bias in causal estimation [20], [21]. This issue becomes particularly pronounced in high-dimensional observational datasets, where many measured variables may have only weak relationships with treatment or outcomes. Recent work has emphasized the role of regularization and representation learning in improving robustness of causal estimators in such settings [22], [23], [24]. However, systematic empirical evaluation of robustness to irrelevant covariates—especially in real-world clinical data—remains limited. In this paper, we address this gap by introducing a controlled irrelevant-feature stress test and examining how treatment effect estimators behave under increasing nuisance variation.

3. Problem Formulation

We formulate individualized treatment effect (ITE) estimation for vasopressor administration in observational ICU data using the potential outcomes framework. This section defines the data structure, identification assumptions, policy evaluation setup, and the latent-factor perspective on covariates that motivates the disentangled representation used by DRI-ITE.

3.1 Potential Outcomes Framework and Observational Data

We consider an observational dataset $\mathcal{D} = \{(x_i, t_i, y_i)\}_{i=1}^{N}$, where each instance (e.g. an ICU patient) has covariates $x_i \in \mathcal{X} \subseteq \mathbb{R}^K$ representing pre-treatment variables, a binary treatment assignment $t_i \in \mathcal{T} = \{0,1\}$indicating whether vasopressors were administered, and an observed outcome $y_i \in \mathcal{Y}$. In our application, y_i denotes a clinically motivated lactate-change endpoint measuring the change in serum lactate levels during the early ICU stay.

Under the potential outcomes framework [1], [2], [8], each patient has two potential outcomes: y_i^0 if untreated and y_i^1 if treated. However, for each individual we only observe the *factual* outcome

$$y_i = y_i^{t_i},$$

while the counterfactual outcome $y_i^{1-t_i}$ remains unobserved. In observational clinical datasets, treatment assignment depends on patient covariates such as physiological measurements, laboratory values, and clinician judgment. Consequently the treatment propensity $P(T = 1 \mid X)$varies across the covariate space, introducing selection bias and making the data fundamentally different from randomized experiments.

We adopt the following standard and widely used assumptions from causal inference.

Stable Unit Treatment Value (SUTVA): There is no interference between units and no hidden versions of treatment, so that each unit's potential outcomes depend only on its own treatment assignment.

Consistency: The observed outcome corresponds to the potential outcome under the received treatment:

$$y_i = y_i^{t_i}.$$

Unconfoundedness: The potential outcomes are conditionally independent of treatment given covariates:

$$(y_i^0, y_i^1) \perp\!\!\!\perp t_i \mid x_i.$$

Overlap: Every treatment has nonzero probability for all covariate values:

$$0 < P(t = 1 \mid x) < 1 \forall\, x \in \mathcal{X}.$$

Unconfoundedness and overlap together imply *strong ignorability* [8]. Under these conditions, the conditional average treatment effect (CATE) is identifiable as

$$\tau(x) = \mathbb{E}[y^1 - y^0 \mid x] = \mathbb{E}[y \mid x, t = 1] - \mathbb{E}[y \mid x, t = 0].$$

At the individual level the treatment effect is defined as

$$\tau_i = y_i^1 - y_i^0.$$

The goal of individualized treatment effect estimation is therefore to learn a predictive model

$$f: \mathcal{X} \times \mathcal{T} \to \mathcal{Y}$$

that estimates counterfactual outcomes $\hat{y}^0(x)$and $\hat{y}^1(x)$.

Although the formulation follows the standard potential outcomes framework, its application to ICU data introduces domain-specific challenges. Treatment assignment reflects clinician-driven decisions under severe and heterogeneous patient conditions, leading to strong confounding and potentially limited overlap between treatment groups. In addition, high-dimensional physiological measurements may include substantial nuisance variation that is weakly related to treatment or outcomes. These characteristics distinguish the present setting from standard benchmark scenarios and motivate the use of representation learning and robustness-focused evaluation.

3.2 Policy Evaluation

Given predicted counterfactual outcomes $\hat{y}^0(x)$ and $\hat{y}^1(x)$from a model f, we define the treatment policy [18], [19]

$$\pi_f(x) = \mathbf{1}\{\hat{y}^1(x) - \hat{y}^0(x) < 0\},$$

which recommends treatment when the predicted individual treatment effect is beneficial, i.e., when the predicted lactate change is negative.

The value of a treatment policy [19] is defined as

$$V(\pi_f) = \mathbb{E}[y^{\pi_f(x)}],$$

representing the expected outcome if the policy were deployed. Because both potential outcomes are not observed simultaneously, we compute a *plug-in estimate* of policy value by substituting model predictions:

$$\hat{V}(\pi_f) = \mathbb{E}[\pi_f(x)\hat{y}^1(x) + (1 - \pi_f(x))\hat{y}^0(x)].$$

This provides a decision-focused diagnostic for comparing models, although it should not be interpreted as an unbiased estimate of real-world clinical utility. To compare policies we evaluate improvement relative to the treat-all policy:

$$\Delta_{\text{all}} = \hat{V}(\pi_f) - \hat{V}(\mathbf{1}).$$

Because the outcome $y_i$represents lactate change where negative values correspond to improvement, more negative values of Δ_{all}indicate larger improvements relative to treating every patient.

3.3 Latent Factor View of Covariates

To motivate the representation-learning approach used in DRI-ITE [7], we adopt a latent-factor perspective on covariates in which different components of $x$may influence treatment assignment and outcomes in different ways. In practice, these factors are not directly observable and may only be approximately separable [15].

We conceptually distinguish four types of latent spaces:

- **Γ (treatment-predictive):** covariate information that primarily influences treatment assignment.
- **Δ (confounding):** variables that influence both treatment assignment and outcomes.
- **Υ (outcome-related):** variables that affect outcomes but are not strong drivers of treatment decisions.
- **Ω (irrelevant or nuisance variation):** covariate information that is weakly related or unrelated to treatment and outcome prediction for the task at hand.

In high-dimensional clinical datasets, irrelevant covariates can inflate variance and degrade treatment effect estimation. Explicitly learning an Ωlatent space provides a mechanism to absorb such variation while encouraging the remaining representations to focus on treatment-relevant
structure.

4 DRI-ITE: Deep Disentanglement with Irrelevant Factors

DRI-ITE was originally introduced in [7]. In this work, we apply the method to observational ICU data and briefly summarize its architecture and objective for completeness.

4.1 Architecture

DRI-ITE learns disentangled representations that separate different roles of covariates in treatment and outcome prediction. The model consists of four encoders producing latent factors $\Gamma(\cdot), \Delta(\cdot), \Upsilon(\cdot), \Omega(\cdot)$representing treatment-predictive, confounding, outcome-related, and irrelevant covariate variation respectively.

Each encoder is implemented as a three-layer MLP with ELU activations and is trained jointly with three downstream components: two outcome regression networks $h_y^0(\cdot), h_y^1(\cdot)$, a treatment classifier $h_c(\cdot)$, and a reconstruction decoder $h_{\text{recon}}(\cdot)$. The regression networks predict potential outcomes and help disentangle Δand Υthrough supervised outcome prediction. The classification network predicts treatment assignment and encourages separation between treatment-predictive (Γ) and confounding (Δ) representations. Since the nuisance latent space Ωreceives no direct supervision signal, the decoder reconstructs the original covariates from all latent factors, encouraging residual covariate information to be absorbed by Ω.

4.2 Objective Function

The overall objective function is:

$$L_{\text{main}} = L_{\text{reg}} + \alpha L_{\text{class}} + \beta L_{\text{disc}} + \gamma L_{\text{recons}} + \lambda L_{\text{orth}} + \mu\, \text{Reg}(h_y^1, h_y^0, h_c, h_{\text{recon}})$$

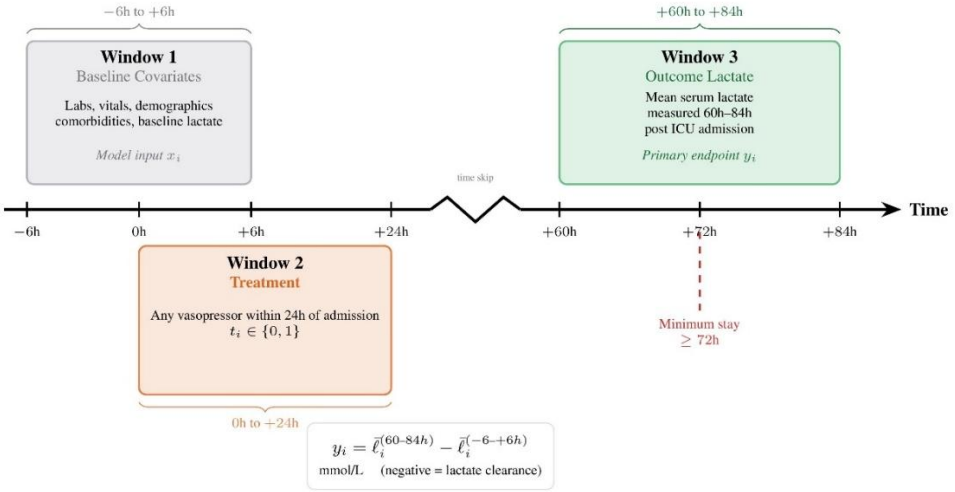

Fig 1 Temporal structure of cohort extraction from MIMIC-III relative to ICU admission. Window 1 (−6h to +6h) defines baseline covariates and reference lactate. Window 2 (0h to +24h) defines treatment, where $t_i = 1$ if any vasopressor is administered. Window 3 (+60h to +84h) defines the outcome lactate. The outcome is the lactate change between the late and baseline windows, where negative values indicate lactate clearance. A minimum ICU stay of 72 hours is required to observe the outcome window.

where Regdenotes L_2weight decay and $\alpha, \beta, \gamma, \lambda, \mu \geq 0$ are hyperparameters controlling the relative contribution of each loss.

Regression loss.

Outcome prediction relies on confounding and outcome-related representations:

$$L_{\text{reg}} = \mathcal{L}[y_i,\ h_y^{t_i}(\Delta(x_i), \Upsilon(x_i))]$$

Two outcome networks are trained, one for each treatment group, following [3], [15]. We use mean squared error (MSE).

Classification loss.

Treatment prediction uses treatment-predictive and confounding factors:

$$L_{\text{class}} = \mathcal{L}\,[t_i,\ h_c(\Gamma(x_i), \Delta(x_i))]$$

The classifier is trained using binary cross-entropy (BCE).

Discrepancy loss.

To reduce selection-induced distribution shift, the model encourages treatment invariance in the Υrepresentation by minimizing a discrepancy between treatment groups:

$$L_{\text{disc}} = \text{disc}[\Upsilon(x_i)\ |_{t_i=0}\ ,\ \Upsilon(x_i)\ |_{t_i=1}]$$

where $\text{disc}(\cdot)$denotes a distribution discrepancy measure (implemented as Wasserstein distance in our experiments), following the discrepancy-based balancing approach of [15].

Reconstruction loss.

The decoder reconstructs the input covariates from all latent factors:

$$L_{\text{recons}} = \mathcal{L}[x_i,\ h_{\text{recon}}(\Gamma(x_i), \Delta(x_i), \Upsilon(x_i), \Omega(x_i))]$$

Because Γ, Δ, and Υare anchored by supervised objectives, the reconstruction task encourages remaining covariate information to be captured by the nuisance latent space Ω.

Orthogonality loss.

To reduce information leakage between latent spaces, we encourage approximate orthogonality among encoder representations. Let $W_\Gamma, W_\Delta, W_\Upsilon$,and $W_\Omega$denote the effective encoder weight matrices, and let $\bar{W}$denote the row-wise average of absolute weights [16]. The orthogonality penalty is

$$\begin{aligned} L_{\text{orth}} = \quad & \bar{W}_\Gamma^\top \bar{W}_\Delta + \bar{W}_\Delta^\top \bar{W}_\Upsilon + \bar{W}_\Upsilon^\top \bar{W}_\Gamma \\ & + \bar{W}_\Omega^\top \bar{W}_\Gamma + \bar{W}_\Omega^\top \bar{W}_\Delta + \bar{W}_\Omega^\top \bar{W}_\Upsilon \end{aligned}$$

The orthogonality constraints involving Ω are particularly important: without them the nuisance latent space may encode information already captured by other latent factors, weakening isolation of irrelevant variation.

4.3 Prediction

At inference time, individualized treatment effects are estimated using only the confounding and outcome-related representations:

$$\hat{\tau}(x_i) = h_y^1(\Delta(x_i), \Upsilon(x_i)) - h_y^0(\Delta(x_i), \Upsilon(x_i))$$

The treatment-predictive representation $\Gamma(x_i)$is excluded from outcome prediction because it primarily

captures treatment assignment signals, while $\Omega(x_i)$represents nuisance covariate variation for the current task and is likewise excluded from outcome prediction.

5 Data and Cohort

We construct an observational cohort from the MIMIC-III critical care database in order to study individualized treatment effects of vasopressor administration. The cohort extraction follows a fixed temporal protocol separating baseline covariates, treatment assignment, and outcome measurement windows relative to ICU admission. Figure 1 illustrates the temporal structure used to define covariates, treatment, and outcome variables.

Table 1 Summary statistics of the constructed vasopressor cohort.

Quantity	Value
Total ICU stays	7,721
Training set size	6,562
Test set size	1,159
Treated proportion	13.8%
Number of covariates (K)	27
Mean outcome E[y_i]	−0.411
Outcome std. dev.	1.337

5.2 MIMIC-III and Cohort Construction

We use the MIMIC-III v1.4 database [6], which contains detailed clinical data for more than 40,000 ICU admissions at Beth Israel Deaconess Medical Center between 2001 and 2012. We construct a cohort of adult ICU stays (≥ 18 years) following a fixed extraction protocol. Only the first ICU stay per patient is included to avoid repeated-measure bias. Patients are required to have lactate measurements in both the baseline window (-6 to $+6$hours around ICU admission) and the follow-up window (60--84 hours), and an ICU stay of at least 72 hours to ensure that the outcome window can be observed. These windows are chosen to capture the initial physiological state and subsequent short-term response while allowing for variability in measurement timing.

Treatment.

The treatment variable $t_i$indicates whether a vasopressor was administered within the first 24 hours of ICU admission:

$$t_i = \begin{cases} 1 & \text{vasopressor administered} \\ 0 & \text{otherwise.} \end{cases}$$

binary indicator reflecting whether vasopressor was administered during the window, without imposing dosage thresholds, to avoid introducing additional modeling assumptions.

Outcome.

The outcome variable $y_i$is defined as the change in serum lactate between a baseline window and a delayed follow-up window:

$$y_i = \bar{\ell}_i^{(60\text{–}84h)} - \bar{\ell}_i^{(-6\text{–}+6h)}.$$

Negative values correspond to lactate reduction and therefore improved metabolic status. The baseline window captures the initial physiological state around ICU admission, while the delayed window reflects short-term response to treatment while reducing sensitivity to measurement noise. Patients with missing lactate measurements in either window are excluded from the analysis.

Covariates.

Each instance is represented by $K = 27$pre-treatment covariates $x_i \in \mathcal{X} \subseteq \mathbb{R}^K$including:

- Vital signs: heart rate, systolic and diastolic blood pressure, respiratory rate, temperature, SpO_2
- Laboratory measurements: baseline lactate, creatinine, bilirubin, sodium, potassium, bicarbonate, hemoglobin, platelet count, white blood cell count, blood urea nitrogen, chloride
- Neurological status: Glasgow Coma Scale (GCS)
- Comorbidities: diabetes, congestive heart failure, COPD, renal failure, liver disease, hypertension, arrhythmia
- Demographics: age and sex

Continuous covariates are standardized using training set statistics and missing covariate values are imputed using median imputation computed from the training data, and the same statistics are applied to the test set.

Dataset Statistics

After cohort construction, the dataset is split into training and test sets using an 85/15 split stratified by treatment assignment.

$$N_{\text{train}} = 6{,}562, N_{\text{test}} = 1{,}159, P(t_i = 1 \mid \text{test}) \approx 0.138.$$

Outcome statistics on the test set are

$$E[y_i] = -0.411 \pm 1.337$$
$$E[y_i \mid t_i = 1] = -0.558 \pm 1.435, E[y_i \mid t_i = 0] = -0.387 \pm 1.319.$$

The crude treated-minus-control difference

$$E[y_i \mid t_i = 1] - E[y_i \mid t_i = 0] \approx -0.172$$

is confounded by clinical treatment selection and should not be interpreted causally. Table 1 summarizes the main characteristics of the resulting cohort.

6 Evaluation Protocol

Because counterfactual outcomes $y_i^{\neg t_i}$ are unobserved in observational data, no single evaluation metric can fully characterize the quality of individualized treatment effect estimates. In particular, commonly used metrics such as PEHE are not available, and average-effect measures alone may fail to capture whether a model produces meaningful treatment decisions. We therefore adopt a multi-faceted evaluation protocol designed to assess complementary aspects of model behavior in a clinically relevant setting.

Specifically, we evaluate models along five dimensions: treatment-effect estimation (via model-implied ATE), decision quality (via policy risk and improvement), heterogeneity (via the distribution of predicted treatment effects), subgroup behavior (via stratified analysis across baseline lactate levels), and robustness (via irrelevant-covariate stress tests). This combination of criteria is intended to capture not only whether a model predicts reasonable average effects, but also whether it induces stable, non-degenerate, and clinically meaningful treatment policies under realistic observational conditions.

6.1 Policy Evaluation

Beyond average treatment effects, we evaluate decision quality using the treat-if-benefit policy. This policy recommends treatment when the predicted effect indicates improved outcome (negative lactate change) [18], [19].

Policy performance is measured using the plug-in policy value $\hat{V}(\pi_f)$ defined in Section 3. Because counterfactual outcomes are not observed, policy value is estimated using model-predicted potential outcomes $\hat{y}^1(x)$ and $\hat{y}^0(x)$.

We report policy improvement relative to the treat-all and treat-none policies:

$$\Delta_{\text{all}} = \hat{V}(\pi_f) - \hat{V}(\mathbf{1}), \Delta_{\text{none}} = \hat{V}(\pi_f) - \hat{V}(\mathbf{0})$$

More negative values indicate larger improvements over the baseline policy. These metrics provide a decision-focused comparison of individualized treatment effect models.

6.2 Model-Implied ATE

We first report the model-implied average treatment effect (ATE) [1], [25]:

$$\widehat{\text{ATE}}_{\text{model}} = \frac{1}{N_{\text{test}}} \sum_{i \in \mathcal{D}_{\text{test}}} (\hat{y}^1(x_i) - \hat{y}^0(x_i))$$

Bootstrap confidence intervals are computed using 500 resamples of the test set to quantify sampling uncertainty.

6.3 Subgroup Analysis

To examine treatment heterogeneity, we stratify test patients by quartiles of baseline lactate measured in the baseline window. For each stratum we compute the mean predicted treatment effect

$$E[\hat{y}^1(x_i) - \hat{y}^0(x_i) \mid \text{stratum}]$$

and the corresponding policy treat rate

$$E[\pi_f(x_i) \mid \text{stratum}]$$

both with bootstrap 95% confidence intervals [9].

6.4 Heterogeneity of Predicted Treatment Effects

To evaluate heterogeneity in predicted treatment effects, we analyze the empirical distribution of $\hat{\tau}(x) = \hat{y}^1(x) - \hat{y}^0(x)$ on the test set. Because true individual treatment effects are not observed, these statistics provide a diagnostic characterization of model-implied heterogeneity rather than a direct measure of estimation accuracy. We report summary statistics including mean(τ), standard deviation, and extreme percentiles (p1, p99), which capture both overall dispersion and tail behavior of predicted treatment effects.

6.5 Irrelevant-Covariate Stress Test

To evaluate robustness to irrelevant covariates, we augment the covariate vector x_i with $|\Omega| \in \{5,10,15,20\}$ synthetic features drawn i.i.d. from $\mathcal{N}(0,1)$. These synthetic features are appended to both the training and test sets. All baseline models are retrained on the augmented datasets [22]. For DRI-ITE, separate checkpoints are trained for each noise level $|\Omega|$. We report policy improvement Δ_{all} and policy treat rate across noise levels. Together, these evaluation criteria assess whether treatment-effect models produce stable and clinically plausible decision policies in observational ICU data where counterfactual outcomes are not observed.

7 Baseline Methods

We compare DRI-ITE against several standard heterogeneous treatment effect estimators widely used in causal machine learning.

S-learner.

A single outcome model is trained to predict y_i from the augmented feature vector (x_i, t_i). Counterfactual outcomes are obtained by intervening on the treatment indicator t_i. We implement the S-learner using both Ridge regression and a Random Forest (RF) regressor [9].

T-learner.

Separate outcome models $\hat{\mu}_0$ and $\hat{\mu}_1$ are trained on the control $(t_i = 0)$ and treated $(t_i = 1)$ subsets,

respectively. Counterfactual outcomes are predicted using the model corresponding to the desired treatment condition. We use Ridge and RF regressors [9].

X-learner.

The X-learner [9] is a two-stage procedure designed for settings with treatment imbalance. First, outcome models $\hat{\mu}_0$and $\hat{\mu}_1$are estimated. Pseudo-treatment effects are then imputed for each unit using the observed outcome and the predicted counterfactual outcome. Separate effect models $\hat{\tau}_0(x)$and $\hat{\tau}_1(x)$are trained on the control and treated groups. The final treatment effect estimate is a propensity-weighted combination:

$$\hat{\tau}(x) = g(x)\,\hat{\tau}_0(x) + (1 - g(x))\,\hat{\tau}_1(x)$$

where $g(x) = P(t = 1 \mid x)$denotes the propensity score. We implement the X-learner using Ridge regression.

Tabel 2 Plug-in policy evaluation under the treat-if-benefit rule. DRI-ITE achieves the largest improvement over the treat-all policy ($\boldsymbol{\Delta_{all}}$) and the lowest policy risk $\boldsymbol{\mathcal{R}_{pol}(\pi_f)}$, while several baselines collapse toward near-degenerate policies (e.g., treat-all). This shows that models with similar ATE can induce substantially different decision policies. Values are reported as mean ± standard deviation over bootstrap resamples; best values are in bold.

Method	$\mathcal{R}_{\text{pol}}(\mathbf{0})$	$\mathcal{R}_{\text{pol}}(\mathbf{1})$	$\mathcal{R}_{\text{pol}}(\boldsymbol{\pi_f})$	Δ_{all}	CI	Δ_{none}	Treat rate
X-learner (Ridge)	-0.381±0.000	-0.586±0.000	-0.659±0.000	-0.073±0.000	[-0.081,-0.066]	-0.279±0.000	0.645±0.000
T-learner (Ridge)	-0.381±0.000	-0.587±0.000	-0.659±0.000	-0.072±0.000	[-0.080,-0.065]	-0.279±0.000	0.645±0.000
T-learner (RF)	-0.381±0.000	-0.561±0.002	-0.596±0.001	-0.035±0.001	[-0.041,0.029]	-0.215±0.002	0.710±0.004
S-learner (RF)	-0.401±0.001	-0.456±0.001	-0.456±0.001	0.000±0.000	[0.000,0.000]	-0.054±0.002	0.876±0.016
S-learner (Ridge)	-0.384±0.000	-0.615±0.000	-0.615±0.000	0.000±0.000	[0.000,0.000]	-0.232±0.000	1.000±0.000
Causal Forest	-0.381±0.000	-0.546±0.008	-0.548±0.008	-0.001±0.000	[-0.002,-0.001]	-0.167±0.008	0.965±0.005
DRCFR	-0.358±0.015	-0.536±0.042	-0.603±0.037	-0.068±0.007	[-0.074,-0.061]	-0.245±0.031	0.600±0.016
RLOCFR	-0.358±0.018	-0.551±0.035	-0.613±0.032	-0.062±0.003	[-0.069,-0.056]	-0.255±0.027	0.621±0.017
DRI-ITE (ours)	-0.372±0.015	-0.539±0.036	**-0.671±0.029**	**-0.132±0.009**	**[-0.147,-0.117]**	-0.299±0.027	0.567±0.014

Causal Forest (CF)

We use the Generalized Random Forest (GRF) implementation [11], which estimates heterogeneous treatment effects using adaptive tree-based partitioning and provides asymptotic inference procedures.

DRCFR.

DRCFR is a disentangled representation-learning approach for treatment effect estimation that separates treatment-related and outcome-related latent factors while encouraging balanced representations across treatment groups [15].

RLOCFR.

RLOCFR extends DRCFR by incorporating regularization mechanisms that promote more structured latent representations and improved robustness to irrelevant covariates [16].

8 Results

We evaluate DRI-ITE on the MIMIC-III vasopressor cohort against standard meta-learners, causal forests, and disentanglement-based baselines. Throughout this section, policy quantities should be interpreted as *plug-in* estimates computed from model-predicted potential outcomes rather than estimates of deployed clinical utility.

8.1 Policy-Based Evaluation

We next evaluate decision quality under the treat-if-benefit rule $\pi_f(x) = \mathbf{1}\{\hat{y}^1(x) - \hat{y}^0(x) < 0\}$. Table 2 reports estimated plug-in policy risks and improvements relative to the treat-all policy. Under this criterion, DRI-ITE achieves the largest improvement over treat-all, with $\Delta_{\text{all}} = -0.132 \pm 0.009$ and CI $[-0.147, -0.117]$. It also attains the lowest policy risk $\mathcal{R}_{\text{pol}}(\pi_f) = -0.671 \pm 0.029$among all methods. The strongest meta-learning baselines, X-learner and T-learner with Ridge regression, achieve $\Delta_{\text{all}} \approx -0.073$and -0.072, respectively. The representation-based baselines also improve over several standard learners, with DRCFR at $-0.068 \pm 0.007$and RLOCFR at -0.062 ± 0.003, but both remain substantially below DRI-ITE. DRI-ITE maintains a non-degenerate treatment recommendation rate of 0.567 ± 0.014, avoiding collapse to trivial treat-all or treat-none policies, whereas several baselines exhibit near-degenerate behavior. S-learner (Ridge) recommends treatment for all patients, yielding no gain over treat-all, while Causal Forest remains highly aggressive with a treatment rate of 0.965 ± 0.005. These results highlight that models with similar ATE magnitudes can induce qualitatively different treatment policies when evaluated from a decision perspective.

8.2 Average Treatment Effect Estimates

Table 3 Model-implied ATE estimates with bootstrap confidence intervals. Most methods yield similar ATE magnitudes, including DRI-ITE, indicating that average effects alone provide limited discrimination between models and motivating policy-based evaluation.

Method	Est.	CI_{low}	CI_{hi}
T-learner (Ridge)	-0.207 ±0.000	-0.234±0.000	-0.182±0.001
T-learner (RF)	-0.180 ±0.002	-0.205±0.003	-0.155±0.003
S-learner (Ridge)	-0.232 ±0.000	-0.232±0.000	-0.232±0.000
S-learner (RF)	-0.054 ±0.002	-0.063±0.002	-0.046±0.002
X-learner (Ridge)	-0.206 ±0.000	-0.234±0.000	-0.181±0.001
Causal Forest	-0.166±0.008	-0.183±0.010	-0.150±0.006
DRCFR	-0.177±0.035	-0.209±0.040	-0.145±0.029
RLO-DRCFR	-0.192±0.029	-0.225±0.033	-0.160±0.023
DRI-ITE	-0.168±0.032	-0.212±0.037	-0.124±0.027

Table 3 reports model-implied ATE estimates together with bootstrap confidence intervals. DRI-ITE estimates $\text{ATE} = -0.168 \pm 0.032$with 95% CI $[-0.212, -0.124]$, which is similar in magnitude to the T- and X-learners, Causal Forest, and the disentanglement baselines. Overall, ATE estimates alone provide only weak separation between methods. Several models yield broadly similar average effects while differing much more substantially in their implied treatment policies, motivating a decision-focused evaluation.

Tabel 4 Policy treatment rates stratified by baseline lactate quartiles. DRI-ITE maintains a stable, non-degenerate treatment rate across severity levels, while several baselines exhibit near-uniform or highly aggressive policies with limited variation across subgroups. This indicates differences in how models adapt treatment decisions to patient severity.

Baseline lactate quartile	S(RF)	S(Ridge)	T(RF)	T(Ridge)	X(Ridge)	CF	DRCFR	RLOCFR	DRI-ITE
Q1 (≤ -0.636)	0.475	1.000	0.528	0.419	0.419	0.996	0.702	0.741	0.606
Q2 (-0.636, -0.297]	0.990	1.000	0.845	0.491	0.491	0.993	0.711	0.752	0.615
Q3 (-0.297, 0.340]	0.997	1.000	0.748	0.748	0.745	0.954	0.753	0.771	0.603
Q4 (> 0.340)	0.975	1.000	0.722	0.931	0.935	0.917	0.703	0.722	0.606

8.3 Subgroup Analysis by Baseline Lactate

To examine how treatment decisions vary with baseline severity, we stratify patients by baseline lactate quartiles and report policy treatment rates in Table 4. DRI-ITE recommends treatment for roughly 60% of patients across quartiles, with only modest variation across severity groups. This pattern is more balanced than the highly aggressive policies learned by several baselines: S-learner (Ridge) recommends treatment for all patients in every quartile, and Causal Forest remains above 0.91 even in the highest-lactate quartile. The representation-based baselines also produce non-degenerate policies, but with higher treatment rates than DRI-ITE across all quartiles. DRCFR and RLOCFR remain in the range of roughly 0.70--0.77, whereas DRI-ITE stays closer to a balanced 0.60. These patterns are consistent with the main policy results: DRI-ITE does not collapse to treat-all while maintaining a stable decision rule across severity strata.

8.4 Heterogeneity of Predicted Treatment Effects

To further characterize the heterogeneity captured by each model, we examine the empirical distribution of predicted individual treatment effects $\hat{\tau}(x) = \hat{y}^1(x) - \hat{y}^0(x)$on the test set. Table 5 summarizes the mean, standard deviation, and extreme percentiles of the predicted treatment-effect distributions. Several baselines exhibit limited variability. In particular, S-learner (Ridge) produces a nearly constant effect for all patients, with essentially

Table 5. Distribution of predicted treatment effects $\hat{\tau}(x)$on the test set. DRI-ITE exhibits the largest spread (Std, p1, p99), indicating richer heterogeneity, while some baselines (e.g., S-learner with Ridge) produce near-constant predictions. This suggests that models differ substantially in the degree of patient-level variation they capture.

Method	Mean $\hat{\tau}$	Std($\hat{\tau}$)	p1	p99
S-learner (RF)	−0.053	0.145	−0.787	0.002
S-learner (Ridge)	−0.232	0.000	−0.232	−0.232
T-learner (RF)	−0.181	0.450	−2.185	0.538
T-learner (Ridge)	−0.207	0.453	−1.841	0.620
X-learner (Ridge)	−0.206	0.456	−1.863	0.623
Causal Forest	−0.181	0.310	−1.476	0.014
DRCFR	−0.180	0.571	−2.754	0.494
RLO-DRCFR	−0.195	0.567	−2.697	0.495
DRI-ITE	−0.172	0.775	−3.572	1.227

zero dispersion, which is fully consistent with its degenerate treat-all policy. S-learner (RF) also exhibits comparatively narrow dispersion. The T- and X-learners produce broader distributions, with standard deviations around 0.45and positive upper tails. The representation-based baselines further increase treatment-effect dispersion. DRCFR and RLO-DRCFR both produce wider distributions than the meta-learners, but DRI-ITE yields the broadest distribution overall, with $\text{std} = 0.775$, $\text{p1} = -3.572$, and $\text{p99} = 1.227$.

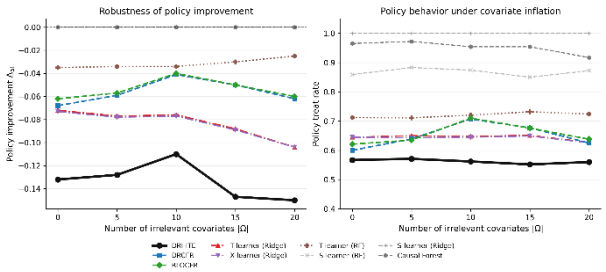

Fig 2 Robustness under irrelevant-feature inflation. Left: policy improvement Δ_{all}as the number of appended nuisance covariates increases (more negative is better). Right: corresponding treatment rate. DRI-ITE maintains the strongest and most stable policy improvement across noise levels, while several baselines degrade or drift toward more aggressive, near-degenerate policies.

This indicates that DRI-ITE captures both strongly beneficial and potentially harmful responses, rather than collapsing toward a near-constant effect estimate. Importantly, this broader predicted heterogeneity does not translate into an extreme treatment policy: DRI-ITE maintains a balanced treatment rate while achieving the largest policy improvement in Table 2. Taken together, these results suggest that the learned disentangled representation captures richer patient-level variation without inducing degenerate or overly aggressive treatment policies.

8.5 Robustness to Irrelevant Covariates

We assess robustness to nuisance variation by appending synthetic Gaussian covariates to the feature set. Table 6 reports policy improvement Δ_{all} and policy treatment rates as irrelevant features are added. Several baselines degrade noticeably under this perturbation. Causal Forest remains close to treat-all

Tabel 6 Robustness under irrelevant-feature inflation. Left: policy improvement Δ_{all}as the number of appended nuisance covariates increases (more negative is better). Right: corresponding treatment rate. DRI-ITE maintains the strongest and most stable policy improvement across noise levels, while several baselines degrade or drift toward more aggressive, near-degenerate policies.

Method	Δ_{all}					Treat rate				
	0	5	10	15	20	0	5	10	15	20
S-learner (RF)	0.000	0.000	0.000	0.000	0.000	0.859	0.883	0.874	0.850	0.872
S-learner (Ridge)	0.000	0.000	0.000	0.000	0.000	1.000	1.000	1.000	1.000	1.000
T-learner (RF)	−0.035	−0.034	−0.034	−0.030	−0.025	0.712	0.711	0.720	0.732	0.724
T-learner (Ridge)	−0.072	−0.077	−0.076	−0.088	−0.104	0.645	0.650	0.648	0.651	0.627
X-learner (Ridge)	−0.073	−0.078	−0.077	−0.089	−0.104	0.645	0.643	0.645	0.649	0.626
Causal Forest	0.000	0.000	0.000	0.000	0.000	0.965	0.972	0.954	0.954	0.917
DRCFR	−0.068	−0.059	−0.041	−0.050	−0.062	0.600	0.639	0.707	0.677	0.625
RLOCFR	−0.062	−0.057	−0.040	−0.050	−0.060	0.621	0.635	0.710	0.676	0.638
DRI-ITE	**−0.132**	**−0.128**	**−0.110**	**−0.147**	**−0.150**	0.567	0.571	0.562	0.552	0.560

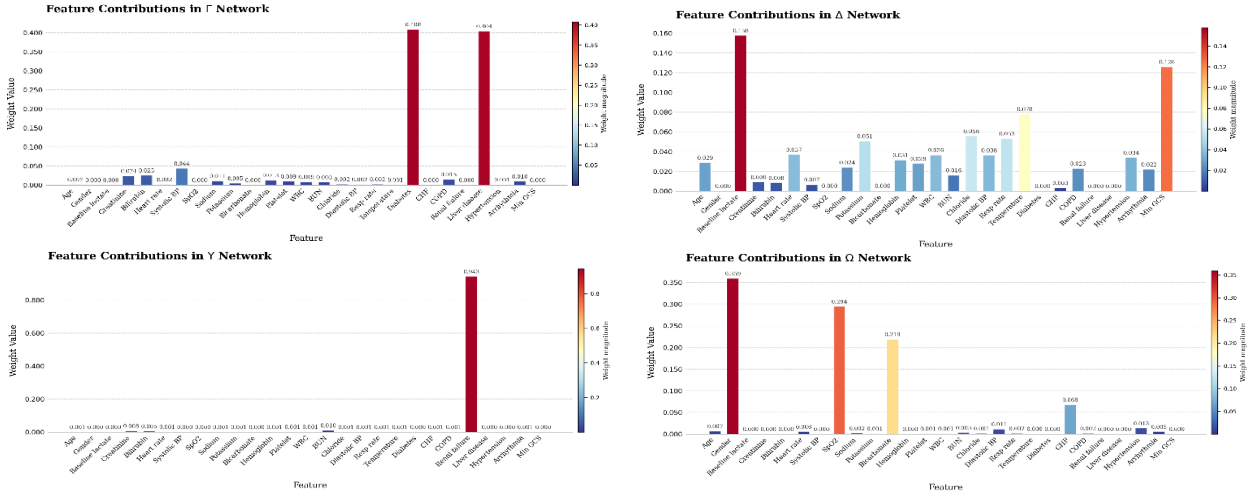

Fig 3 Feature-contribution profiles for the four latent factors. Higher values indicate stronger reliance on a covariate within each factor. The confounding factor **Δ**concentrates on acute severity markers, **Υ**captures outcome-related variation (e.g., renal failure), **Γ**reflects treatment-predictive structure, and **Ω**absorbs nuisance variation. Overall, the separation illustrates how the model organizes covariates into distinct functional roles in the learned representation.

throughout, and S-learner (Ridge) remains fully degenerate at every noise level. The Ridge-based meta-learners are more stable but continue to achieve substantially smaller policy improvements than DRI-ITE. The representation-based baselines remain non-degenerate under feature inflation, but their improvements are consistently weaker than those of DRI-ITE. DRCFR ranges from −0.068to −0.041and RLOCFR from −0.062to −0.040, while both shift toward more aggressive treatment policies as irrelevant covariates are added. In contrast, DRI-ITE maintains strong policy improvement at every noise level, ranging from −0.110to −0.150, while its treatment rate remains tightly controlled between 0.552 and 0.571. These results suggest that the disentanglement objective used by DRI-ITE helps provide a more robust separation between

treatment-relevant structure and nuisance variation, reducing the tendency of the policy to drift toward indiscriminate treatment when irrelevant covariates are introduced. Figure 2 complements Table 6 by visualizing these trends. The left panel shows that DRI-ITE consistently outperforms competing methods across all noise levels, whereas other methods deteriorate more substantially as | Ω |increases. The right panel shows that DRI-ITE preserves a stable and balanced treatment rate, while several baselines drift toward more aggressive treatment behavior.

8.6 Role Identification via Disentangled Representations

Finally, we examine the learned feature-role decomposition. Figure 3 visualizes absolute feature-contribution profiles for the four latent factors. The confounding factor Δis primarily associated with acute physiological indicators including baseline lactate, multiple vital signs, metabolic laboratory measurements, and neurological status. This is clinically plausible, as such variables can influence both vasopressor administration and subsequent lactate trajectory. The adjustment factor ϒis strongly influenced by renal failure, suggesting that renal dysfunction carries outcome-related information that is not primarily routed through the treatment-predictive factor. The factor denoted Γis best interpreted as a *model-implied treatment-predictive latent space* rather than a formally validated set of instrumental variables. It is driven by features such as diabetes, COPD, liver disease, and systolic blood pressure, indicating that the learned treatment-assignment pathway reflects both chronic comorbidity burden and hemodynamic state. The nuisance factor Ωcaptures variables such as gender, SpO_2, bicarbonate, and CHF that exhibit comparatively lower contribution to the supervised prediction objectives under the imposed factorization. Overall, the separation between factors is broadly consistent with the intended behavior of the orthogonality and reconstruction objectives.
This separation improves interpretability and robustness by isolating nuisance variation and distinguishing confounding from outcome-related factors, ensuring that predicted effects reflect clinically meaningful signals. At the same time, these assignments should be interpreted as *model-implied* rather than externally validated causal categories.

Summary: Across these complementary diagnostics, DRI-ITE produces the strongest policy improvements while maintaining a balanced treatment rate and remaining comparatively robust to irrelevant feature inflation. Although several baselines yield broadly similar average treatment effects, policy-based evaluation, heterogeneity diagnostics, and robustness analyses reveal substantial differences in the quality and stability of individualized treatment decisions.

9. Discussion

Why ATE comparisons are insufficient

Most methods produce ATE estimates of similar magnitude , yet their policy behavior differs sharply . In particular, S-learner (Ridge) attains a competitive model-implied ATE while recommending treatment for every patient, yielding no policy improvement over treat-all. This disconnect illustrates why evaluating individualized treatment effect estimators solely through average-effect recovery can be misleading in observational clinical settings.

Degenerate policies as a failure mode

Near-universal treatment emerges as a systematic failure mode for several baselines, especially under irrelevant feature inflation. Such models may still report reasonable average effects, but they fail to produce actionable individualization. Explicitly reporting policy treatment rates and subgroup behavior is therefore essential for diagnosing whether a model is genuinely personalizing or merely reproducing an almost constant treatment recommendation.

Why DRI-ITE is more robust

A possible explanation for DRI-ITE's robustness is that the orthogonality and reconstruction objectives discourage nuisance information from entering the latent spaces used for treatment effect prediction. The orthogonality loss L_{orth}limits overlap between learned latent spaces, while the reconstruction loss L_{recons}provides an explicit sink for irrelevant variation. Together, these mechanisms may help preserve meaningful heterogeneity in predicted treatment effects $\hat{\tau}_i$under irrelevant-feature inflation.

Clinical implications

The DRI-ITE policy treats roughly half of the patients, and this recommendation rate remains moderate across baseline lactate strata. This suggests that the model identifies a sizable subgroup that may not benefit from vasopressor administration under the current observational treatment pattern. At the same time, these findings are derived entirely from observational data and plug-in counterfactual predictions; they should be viewed as hypothesis-generating rather than deployment-ready.

Limitations

- *Unmeasured confounding*: The unconfoundedness assumption is untestable, and unrecorded severity markers or clinical decision factors may bias all estimates.
- *Plug-in policy evaluation*: Policy value $\mathcal{R}_{\text{pol}}(\pi_f)$is computed from model-predicted counterfactual outcomes rather than observed outcomes under deployment. Consequently, it

reflects model behavior rather than realized clinical utility.

- *Limited support and extrapolation*: Despite substantial overall propensity-score overlap, the learned policy recommends treatment more frequently than observed, including in low-propensity regions. This reflects broadly negative predicted treatment effects but also indicates reliance on extrapolation, making estimates in such regions inherently more uncertain and requiring cautious interpretation.
- *Outcome proxy*: Lactate change is an intermediate physiological endpoint and may not fully capture longer-term clinical outcomes such as mortality, organ recovery, or length of ICU stay.
- *External validity*: MIMIC-III is derived from a single tertiary-care center in the United States, which may limit generalizability to other hospitals or patient populations.
- *Role interpretation*: Feature-role assignments in the disentangled representation are model-implied and should not be interpreted as externally validated causal categories.

Together, these results suggest that explicitly separating irrelevant covariate variation may improve the stability and interpretability of individualized treatment policies derived from high-dimensional observational clinical data.

10 Conclusion

We presented a policy-focused evaluation framework for individualized treatment effect (ITE) models on MIMIC-III vasopressor data, integrating model-implied ATE estimation, plug-in policy risk evaluation, subgroup analysis, and an irrelevant-covariate stress test. Applying this framework, we showed that standard meta-learners and causal forests, while competitive in ATE recovery, can approach near-degenerate treat-all policies and lose robustness as Ω increases. DRI-ITE, by explicitly disentangling treatment-predictive (Γ), confounding (Δ), outcome-related (Υ), and irrelevant (Ω) covariate variation through reconstruction and orthogonality objectives, maintains heterogeneous treatment effect predictions and achieves larger policy improvements across experimental conditions.

References

[1] D. B. Rubin, "Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies," *J. Educ. Psychol.*, vol. 66, no. 5, pp. 688–701, 1974.

[2] J. Neyman, "On the Application of Probability Theory to Agricultural Experiments," *Stat. Sci.*, vol. 5, no. 4, pp. 465–480, 1990.

[3] U. Shalit, F. D. Johansson, and D. Sontag, "Estimating Individual Treatment Effect: Generalization Bounds and Algorithms," in *Proceedings of the 34th International Conference on Machine Learning*, 2017.

[4] J. M. Robins, A. Rotnitzky, and L. P. Zhao, "Estimation of Regression Coefficients When Some Regressors Are Not Always Observed," *J. Am. Stat. Assoc.*, vol. 89, no. 427, pp. 846–866, 1994.

[5] H. Bang and J. M. Robins, "Doubly Robust Estimation in Missing Data and Causal Inference Models," *Biometrics*, vol. 61, no. 4, pp. 962–973, 2005.

[6] A. Johnson, T. Pollard, and R. Mark, "MIMIC-III Clinical Database," *PhysioNet*, Sep. 2016, doi: 10.13026/C2XW26.

[7] A. S. Khan, E. Schaffernicht, and J. A. Stork, "On the Effects of Irrelevant Variables in Treatment Effect Estimation with Deep Disentanglement," in *European Conference on Artificial Intelligence*, 2024.

[8] P. R. Rosenbaum and D. B. Rubin, "The Central Role of the Propensity Score in Observational Studies for Causal Effects," *Biometrika*, vol. 70, no. 1, pp. 41–55, 1983.

[9] S. R. Künzel, J. S. Sekhon, P. J. Bickel, and B. Yu, "Metalearners for Estimating Heterogeneous Treatment Effects Using Machine Learning," *Proc. Natl. Acad. Sci.*, vol. 116, no. 10, pp. 4156–4165, 2019.

[10] S. Athey and G. W. Imbens, "Recursive partitioning for heterogeneous causal effects," *Proc. Natl. Acad. Sci.*, vol. 113, no. 27, pp. 7353–7360, 2016.

[11] S. Wager and S. Athey, "Estimation and Inference of Heterogeneous Treatment Effects Using Random Forests," *J. Am. Stat. Assoc.*, vol. 113, no. 523, pp. 1228–1242, 2018.

[12] F. Johansson, U. Shalit, and D. Sontag, "Learning Representations for Counterfactual Inference," in *Proceedings of the 33rd International Conference on Machine Learning*, 2016.

[13] C. Louizos, U. Shalit, J. M. Mooij, D. Sontag, R. Zemel, and M. Welling, "Causal Effect Inference with Deep Latent-Variable Models," *Adv. Neural Inf. Process. Syst.*, vol. 30, 2017.

[14] C. Shi, D. M. Blei, and V. Veitch, "Adapting Neural Networks for the Estimation of Treatment Effects," in *Advances in Neural Information Processing Systems*, 2019.

[15] N. Hassanpour and R. Greiner, "Learning Disentangled Representations for Counterfactual Regression," in *International Conference on Learning Representations*, 2020.

[16] M. Cheng, X. Liao, Q. Liu, B. Ma, J. Xu, and B. Zheng, "Learning Disentangled Representations for Counterfactual Regression via Mutual Information Minimization," in *ACM SIGIR Conference on Research and Development in Information Retrieval*, 2022.

[17] M. Dudík, J. Langford, and L. Li, "Doubly robust policy evaluation and learning," in *Proceedings of the 28th International Conference on Machine Learning*, 2011, pp. 1097–1104.

[18] T. Kitagawa and A. Tetenov, "Who Should Be Treated? Empirical Welfare Maximization Methods for Treatment Choice," *Econometrica*, vol. 86, no. 2, pp. 591–616, 2018.

[19] S. Athey and S. Wager, "Policy Learning with Observational Data," *Econometrica*, vol. 89, no. 1, pp. 133–161, 2021.

[20] J. Pearl, "On a Class of Bias-Amplifying Covariates that Endanger Effect Estimates," in *Proceedings of the 26th Conference on Uncertainty in Artificial Intelligence*, 2010.

[21] C. Cinelli and C. Hazlett, "Making sense of sensitivity: Extending omitted variable bias," *J. R. Stat. Soc. Ser. B Stat. Methodol.*, vol. 82, no. 1, pp. 39–67, 2020.

[22] P. R. Hahn, J. S. Murray, and C. M. Carvalho, "Bayesian Regression Tree Models for Causal Inference: Regularization, Confounding, and Heterogeneous Effects," *Bayesian Anal.*, vol. 15, no. 3, pp. 965–1056, 2020.

[23] V. Chernozhukov *et al.*, "Double/debiased machine learning for treatment and structural parameters," *Econom. J.*, vol. 21, no. 1, pp. C1–C68, 2018.

[24] M. J. Vowels, N. C. Camgoz, and R. S. Sherratt, "Targeted VAE: Variational and Targeted Learning for Causal Inference," in *IEEE International Conference on Data Mining Workshops*, 2021.

[25] G. W. Imbens and D. B. Rubin, *Causal Inference for Statistics, Social, and Biomedical Sciences*. Cambridge University Press, 2015.