CONFOUNDER DETECTION VIA TREATMENT INTENT: A NEW OBSERVATIONAL STUDY DESIGN

Anonymous authors
Paper under double-blind review

ABSTRACT

Understanding the effects of interventions is central to scientific progress, with randomized controlled trials (RCTs) regarded as the gold standard for causal inference in many applied fields. However, RCTs are costly, time-consuming, and often constrained by ethical or practical limitations, motivating the need for causal methods able to draw conclusions from observational data. While such data is collected at ever larger scale, making its use for causal inference is often hindered by the fact that not all variables affecting treatment allocation and the outcome are observed – an issue known as unobserved confounding. In this paper, we introduce a new study design called *confounder detection via treatment intent*. The idea is to query a human expert who makes treatment decisions, and ask them to compare pairs of units proposed by a principled matching strategy, with the goal of eliciting unobserved variables that explain why treatment decisions differ. We provide a theoretical basis for such a procedure, ascertaining conditions under which such a study design may elicit unobserved confounders. Building on this newly established foundations, we study treatment effects of interventions in the intensive care unit (ICU). First, we show empirical evidence strongly indicating that electronic health records (EHRs) collected in ICUs are subject to unobserved confounding. By using clinical text notes as a proxy for physicians' knowledge and leveraging natural language processing, we provide a proof of concept for our methodology in a semi-synthetic environment with a known ground truth.

1 INTRODUCTION

Observational data is collected at ever larger scale across empirical sciences, offering a valuable and inexpensive resource for answering scientific questions (Hersh et al., 2013; Jha et al., 2009). This abundance of data also creates an opportunity to draw causal conclusions without resorting to costly experimentation. Such inference, however, does not come for free: moving across the layers of Pearl's Causal Hierarchy (PCH) (Bareinboim et al., 2022) requires appropriate causal assumptions. We focus on the classical setting of treatment effect estimation, where the goal is to infer an interventional (Layer 2) quantity from observational data (Layer 1). The treatment is denoted by X, outcome by Y, and a vector of confounders by Z, where Z causally precedes (X, Y). One of the most common causal assumptions used for identifying treatment effects in such a setting is the back-door criterion (Pearl, 2000), which permits adjustment-based identification of interventional quantities from observational data (Pearl, 2000; Bareinboim, 2025).

The assumptions required for back-door admissibility, however, need not always hold true. For instance, when some of the common causes of X and Y remain unobserved – a situation referred to as unobserved confounding – back-door admissibility does not hold. Such confounding is arguably the most common roadblock faced by causal analysts working with observational data, appearing across empirical sciences, ranging from medicine and epidemiology to economics and the social sciences. Importantly, unobserved confounding can rarely be ruled out from the data alone. In this paper, we use an applied example to illustrate our methodology, specifically in the context of treatment effect estimation in the intensive care unit (ICU), where interventions are allocated by physicians attending to patients.

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.

Example 1 (Mechanical Ventilation from EHR data)**.** *Mechanical ventilation* (X) *is one of the key treatments for ICU patients, used to support or improve oxygenation, while in-hospital mortality* (Y) *is a typical outcome of interest. Adjusting for a set of observed covariates* Z *(including age, sex, SOFA score, and physiological signals; see Sec. 2), we estimate the effect of the treatment on the treated,*

$$\text{ETT} := \mathbb{E}[Y_{x_1} - Y_{x_0} \mid X = x_1], \tag{1}$$

across three large ICU databases (MIMIC-III (Johnson et al., 2016), AUMCdb (Thoral et al., 2021), SICdb (Rodemund et al., 2023)). Fig. 1 shows the estimated $\widehat{\text{ETT}}^{bd}$ *values, which are strongly above zero in every dataset. Under the assumption of no UCs, one would conclude that mechanical ventilation increases mortality among the patients who actually received it. While it is possible that some patients in the broader population may be harmed by this treatment, domain experts would strongly reject such a finding of harm in the treated population, who are often ventilated for a reason. The estimate is therefore a near-certain signature of UCs (see extended discussion in App. D).* □

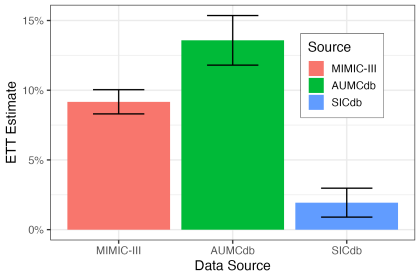


Figure 1: Estimated ETT of mechanical ventilation on in-hospital mortality across three ICU databases.

How does one proceed in the presence of such unobserved confounding? The first broad alternative is to collect experimental data (for instance, for the ETT in Eq. 1 one could apply *counterfactual randomization* (Bareinboim et al., 2015)), which would control for UCs by design. However, this alternative does not come without its own difficulties: carrying out trials requires significant resources in terms of cost, study design, preparation, patient recruitment, and conduct, and ethical or logistical constraints may further limit their feasibility – or rule them out entirely (Harhay et al., 2014; Schulz, 1996). The second alternative is to leverage domain knowledge to identify additional covariates that should be measured, and to collect data on them in a future iteration of the study. Eliciting such domain knowledge directly, however, may be non-trivial: experts are rarely trained to reason in such terms, and asking an ICU physician to enumerate the unmeasured causes of the mechanical-ventilation-to-mortality relationship is a daunting exercise. Interestingly, systematic ways for eliciting unmeasured confounders have received little investigation.

In this paper, we propose an observational study design that attempts to systematically elicit candidate unobserved confounders from a decision-maker (DM) who assigns treatment. The key observation is that any U that confounds the $X \to Y$ relationship must be available to the DM at the time of the decision – otherwise it could not affect X. Suppose then that (i) we may interact with the DM, and (ii) the DM has the cognitive ability to compare different sample pairs. Under these two assumptions, rather than asking them to list unmeasured confounders in the abstract, we may ask them to compare *pairs* of units (i, j) with $X_i = 1$ (treated) and $X_j = 0$ (untreated), with the intention of eliciting why the treatment decision differed.

Example 1 (Continued – Comparing Ventilated Pairs)**.** *Let* $Z^{(1)}$ *denote the SOFA score, a standard ICU severity score ranging from 0 to 24, with higher values corresponding to greater illness severity. Let* $U^{(1)}$ *be the patient's difficulty breathing perceived by the physician, which is not recorded in the EHR. We consider three candidate pair proposals, where in each a treated patient* i *(*$X_i = 1$*) and an untreated patient* j *(*$X_j = 0$*) are compared:*

(1) Patients i *and* j *have the same SOFA score,* $Z_i^{(1)} = Z_j^{(1)} = 5$*, meaning neither is sicker along the observed* $Z^{(1)}$*. If asked for a reason for different treatment between* i *and* j*, a physician who made the decision with access to* $(Z^{(1)}, U^{(1)})$ *would most likely not cite* $Z^{(1)}$ *as the explanation of their decision. If* $U_i^{(1)} = 0$*, while* $U_j^{(1)} = 1$*, variable* $U^{(1)}$ *would be a more likely explanation.*

(2) The treated patient i *has* $Z_i^{(1)} = 5$ *while the untreated patient* j *has* $Z_j^{(1)} = 10$*, meaning patient* j *is sicker along* $Z^{(1)}$*. In such a setting, the DM would again likely not cite* $Z^{(1)}$ *as the explanation for their treatment decision.*

(3) Finally, suppose that $Z^{(2)}$ (sex) is available, and we have

$$Z_i^{(1)} = 5,\ Z_i^{(2)} = 1,\ U_i^{(1)} = 0, \tag{2}$$

$$Z_j^{(1)} = 5,\ Z_j^{(2)} = 0,\ U_j^{(1)} = 1. \tag{3}$$

In such a setting, determining $U^{(1)}$ as the explanation of differing treatment would be made more difficult compared to Case 1, due to the additional existence of variable $Z^{(2)}$ along which the patients differ, which may deter the DM. □

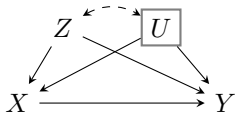


Figure 2: Causal diagram.

The above example illustrates some of the key principles behind our framework. Case (1) argues that if patients match on Z, the DM is forced to name a variable outside Z to explain the treatment difference – e.g., $U^{(1)}$. Case (2) leverages a natural monotonicity consideration of this setting: if the decisions are monotone with respect to $Z^{(1)}$ (which is the case with the SOFA score in ICU data), then seeing the sicker patient untreated would render $Z^{(1)}$ an unlikely explanation – again pointing to $U^{(1)}$. Case (3) shows that comparisons become harder whenever patients differ along observed covariates that are not clearly monotonically related to the outcome (or when monotonic covariates point to different illness levels), since such differences invite explanations related to Z, and do not reveal UCs. These observations raise the central question of this paper: are there systematic, principled ways to propose patient pairs to a domain expert so that the elicited explanations are likely to correspond to genuine unobserved confounders? Specifically, our contributions in the paper are as follows:

(i) We formalize a new observational study design, *confounder detection via treatment intent* (CDTI). We further develop a framework for analyzing such a study design, comprising a matching strategy $\mathcal{M}$ (how sample pairs are proposed to a human annotator) and an extraction strategy ε (how the decision-maker reasons over a pair).

(ii) We introduce specific matching strategies – Z-matching, π-matching, and Z-dominance – and provide theoretical support for them by establishing stochastic-dominance results that characterize when each strategy yields pair distributions informative about U (Thm. 1). We then establish further results on dominance between strategies (Thms. 2 and 3).

(iii) We illustrate our study design based on ICU data. First, we provide evidence that EHR-based estimates of the effect of common ICU interventions are strongly confounded (Ex. 1). Then, we further validate our study design based on semi-synthetic MIMIC-III data.

Related work. The ideas in this paper are related to counterfactual randomization (Bareinboim et al., 2015), which acknowledges that the DM has access to information about U that is not captured in the recorded covariates. Our work, however, uses this observation with a different purpose, namely for eliciting UCs through structured queries. Further, our work is also related to research on combining observational and experimental data to sharpen causal conclusions (Athey et al., 2020; Rosenman et al., 2023; Lee et al., 2020); in contrast, however, our study design remains purely observational and does not modify treatment decisions; in other words, it does not require experimentation on the treatment variable X. Finally, our work is related to the notion of thick description (Rosenbaum & Silber, 2001), which is an informal method for comparing treated-untreated pairs; in contrast, our work offers a systematic approach for eliciting confounders through specific pair-matching strategies.

2 Confounder Detection via Treatment Intent

We now describe our theoretical setting. We consider a treatment variable X (e.g., mechanical ventilation in ICU), outcome variable Y (in-hospital mortality), a set of observed confounders Z (SOFA score, age, sex), and a set of unobserved confounders U (e.g., presence of hemothorax or other lung complications) as in Fig. 2. We have data on (Z, X, Y). Our goal is to recover some of the unobserved confounders U, where we assume access to some kind of proxy information for U. In practice, the idea is that we may be able to query the DM controlling X, and ask them to compare two units: unit i who was treated $(X_i = 1)$ and unit j who was not $(X_j = 0)$. If it happens that unit i is smaller than unit j on each observed confounder, $Z_i^{(k)} < Z_j^{(k)}$ for all k, then the DM may

be able to identify some of the U variables: e.g., it may happen that $U_i^{(l)} > U_j^{(l)}$, so that for the unobserved confounder $U^{(l)}$ unit i is larger, which actually explains the decision. Theoretically, the first important part of our framework is a *matching strategy* $\mathcal{M}$. $\mathcal{M}$ serves for generating *proposals* that the DM can then compare; e.g., $\mathcal{M}$ may generate pairs (i, j) that are sent to the human annotator. We consider several matching strategies:

(i) $\mathcal{M}^{Z\text{-match}}$ picks pairs with $X_i = 1, X_j = 0$ and $Z_i = Z_j$,

(ii) $\mathcal{M}^{\pi\text{-match}}$ picks pairs with $X_i = 1, X_j = 0$ and $\pi(Z_i) = \pi(Z_j)$ where $\pi(Z) = P(X = 1 \mid Z)$,

(iii) $\mathcal{M}^{Z\text{-dom}}$ picks pairs with $X_i = 1, X_j = 0$ and $Z_i \leq Z_j$ coordinatewise,

(iv) $\mathcal{M}^{\text{marg}}$ picks pairs (i, j) with just $X_i = 1, X_j = 0$; this strategy serves as a baseline.

We begin by immediately stating one of our main theoretical results, which anchors our framework and formalizes the intuition from Sec. 1 (all proofs are provided in App. A):

Theorem 1 (Stochastic dominance of matching strategies)**.** *Under appropriate assumptions $\mathcal{A}^{(1)}, \mathcal{A}^{(2)}, \mathcal{A}^{(3)}$, respectively, for each strategy $\mathcal{M} \in \{\mathcal{M}^{Z\text{-match}}, \mathcal{M}^{\pi\text{-match}}, \mathcal{M}^{Z\text{-dom}}\}$, we have:*

$$P(U \mid Z = z, X = 1) \succeq_{st} P(U \mid Z = z, X = 0) \tag{4}$$

$$P(U \mid \pi = p, X = 1) \succeq_{st} P(U \mid \pi = p, X = 0) \tag{5}$$

$$P(U \mid Z = z', X = 1) \succeq_{st} P(U \mid Z = z, X = 0) \text{ if } z' < z, \tag{6}$$

where the $\succeq_{st}$ denotes the multivariate stochastic order, $A \succeq_{st} B \implies \mathbb{E}[\phi(A)] \geq \mathbb{E}[\phi(B)]$ for every coordinatewise non-decreasing $\phi : \mathbb{R}^d \to \mathbb{R}$.

Intuitively, the theorem shows that conditioning on $X = 1$ vs. $X = 0$ (for different additional contexts $C = c$) results in a *probabilistically larger* U for the treated unit, which justifies why a matching strategy may work. With this target established, we now build towards Thm. 1 sequentially, unpacking the necessary technical conditions and providing the intuition for each matching strategy.

2.1 Z-MATCHING

We start with the Z-matching strategy, and first consider the basic case of univariate U:

Proposition 1 (Z-matching via MLR)**.** *Assume $U \in \mathbb{R}$ and that $P(X = 1 \mid Z = z, U = u)$ is non-decreasing in u for every z. Then*

$$P(U \mid Z = z, X = 1) \succeq_{st} P(U \mid Z = z, X = 0). \tag{7}$$

Our first result illustrates that if U has a monotonic effect on the treatment X for each $Z = z$, then conditioning on $X = 1$ vs. $X = 0$ (for a fixed $Z = z$) results in a probabilistically larger U. In the context of our running example (see Ex. 1), this means that for two patients with the same SOFA score $Z^{(1)} = 5$, the one treated ($X = 1$) has a probabilistically larger difficulty breathing $U^{(1)}$ compared to the untreated one ($X = 0$). We note that the monotonic effect assumption is plausible in many health contexts, and many of the covariates used in ICU to assess patient state satisfy this property. Throughout the paper, the best way to think about the X-mechanism is as a *mixture decision-maker*, i.e., a probabilistic average over the individual decision-makers: $P(x \mid z, u) = \mathbb{E}_G[P_g(x \mid z, u)]$. Any condition involving the X-mechanism is required only for this "average" over individual DMs, which is weaker than requiring them for each DM (e.g., individual $P_g(x \mid z, u)$ may violate a condition while the mixture does not). In summary, the above result provides the basis for why it makes sense to send Z-matched treated-untreated unit pairs for comparison to a human annotator. We next move onto the case of multivariate U, and prove a key stochastic dominance result for it:

Proposition 2 (Z-matching, multivariate U)**.** *Assume $U \in \mathbb{R}^d$ and:*

(i) $P(X = 1 \mid Z = z, U = u)$ is non-decreasing in each coordinate of u, for every z;

(ii) $P(U \mid Z = z)$ is log-supermodular, i.e., $P(u \mid z)\, P(u' \mid z) \leq P(u \wedge u' \mid z)\, P(u \vee u' \mid z)$ for all u, u', where $\wedge, \vee$ denote coordinatewise minima and maxima. For twice-differentiable densities, this is equivalent to $\frac{\partial^2 \log P(u|z)}{\partial u^{(i)} \partial u^{(j)}} \geq 0$; Then

$$P(U \mid Z = z, X = 1) \succeq_{st} P(U \mid Z = z, X = 0). \tag{8}$$

The above result shows that stochastic dominance of Z-matching for a univariate U can be extended to a multivariate U, provided that the conditional $P(U \mid Z = z)$ is *log-supermodular*. For twice-differentiable functions, log-supermodularity is equivalent to saying that cross-component second order derivatives $\frac{\partial}{\partial u^{(i)} \partial u^{(j)}} \log P(U \mid Z = z)$ are greater than 0. For instance, in the Gaussian setting, log-supermodularity requires the off-diagonal precision matrix entries to be non-positive, $(\Sigma^{-1})_{kl} \leq 0$ (see Ex. 2 of App. B). In the context of our ICU example, this means that among patients with identical observed covariates $Z = z$, the unobserved severity markers $U^{(1)}, U^{(2)}, \ldots$ (e.g., perceived difficulty breathing, hemodynamic instability not captured in Z) are *positively dependent*: a patient unusually severe along one unobserved axis is, on average, also more severe along the others. This rules out scenarios in which the unobserved factors *trade off* against one another at fixed $Z = z$, but accommodates the more common clinical reality that severity tends to cluster across organ systems. Ex. 3 of App. B shows why log-supermodularity in Prop. 2 is necessary.

2.2 π-MATCHING

We next move onto the π-matching strategy, showing that it inherits the result from the Z-matching case, just by invoking the propensity's covariate balancing property $X \perp\!\!\!\perp Z \mid \pi(Z)$. Specifically, we have the following key result:

Proposition 3 (π-matching, multivariate U)**.** *Under the monotonicity and log-supermodularity assumptions of Prop. 2, for every $p \in (0, 1)$,*

$$P(U \mid \pi(Z) = p, X = 1) \succeq_{st} P(U \mid \pi(Z) = p, X = 0). \tag{9}$$

The covariate balancing property ensures that within a level set $\{Z : \pi(Z) = p\}$, the distribution of Z is the same across treated and untreated groups; hence the only systematic difference between the two groups is the one induced by U, which is exactly the shift quantified by Prop. 2. More generally, the same argument applies to any *balancing score* $b(Z)$ satisfying $X \perp\!\!\!\perp Z \mid b(Z)$, of which Z itself and $\pi(Z)$ are the two extremes: Z being the multi-dimensional, most granular score, and π being the one dimensional, coarsest score. In the context of our ICU example, this result means that pairs of patients with the same propensity for mechanical ventilation – but where one was actually ventilated and the other was not – exhibit a systematic shift in unobserved severity in the expected direction, under the same assumptions required for Z-matching. This is the property that makes π-matching practically attractive: it admits exact matches in a one-dimensional score even when Z is high-dimensional, while preserving the structural guarantee of Z-matching.

2.3 Z-DOMINANCE

The previous two results exploit conditional X*-dominance*: conditioning on $X = 1$ vs. $X = 0$ at fixed $Z = z$ (or $\pi(Z) = p$) induces a specific shift in U. In practice, however, exact matching on $Z = z$ may be difficult, especially when Z is high-dimensional. Suppose instead we are able to find two units with $Z_i = z'$, $Z_j = z$, where $z' < z$ coordinatewise. Intuitively, the treated unit i is smaller along each observed dimension, which should make it more likely that the reason for treatment lies in the hidden variables U. This intuition relies on chaining two stochastic shifts. The first is the Z-matching shift of Prop. 2, comparing the two treatment groups at the same Z. The second is a new shift, comparing two values of Z at a fixed treatment level $X = x$, called Z*-dominance*:

$$P(U \mid Z = z', X = x) \succeq_{st} P(U \mid Z = z, X = x) \quad \text{for } z' < z. \tag{10}$$

The reasoning behind Eq. 10 is that, among (un)treated units, smaller observed z' means the propensity to be (not) treated must have been compensated for along some other axis – namely U. Chaining either shift with Z-matched X-dominance gives the desired result in Eq. 6. We now identify when Eq. 10 holds, captured in the following result:

Proposition 4 (Z-dominance, multivariate U)**.** *Assume conditions of Prop. 2 and:*

(iii) For some $x \in \{0, 1\}$, defining $h^{(x)}(z, u) := P(X = x \mid z, u)\, P(u \mid z)$, the function $\log h^{(x)}$ is twice differentiable and strictly positive, with

$$\frac{\partial^2 \log h^{(x)}}{\partial u^{(j)}\, \partial u^{(k)}}(z, u) \geq 0 \quad \textit{for all } j \neq k, \qquad \frac{\partial^2 \log h^{(x)}}{\partial z^{(l)}\, \partial u^{(j)}}(z, u) \leq 0 \quad \textit{for all } l, j. \tag{11}$$

Then, for $z' < z$, $P(U \mid Z = z', X = 1) \succeq_{st} P(U \mid Z = z, X = 0)$.

Interpreting condition (iii): collider vs. bidirected. The inequalities in (iii) become more transparent once $\log h^{(x)}$ is split into the contributions from the two channels through which Z and U interact. Writing

$$\log h^{(x)}(z,u) \;=\; \underbrace{\log P(X = x \mid z, u)}_{T_C \text{ (collider channel: } Z \to X \leftarrow U)} \;+\; \underbrace{\log P(u \mid z)}_{T_B \text{ (bidirected channel: } Z \dashleftarrow\dashrightarrow U)}, \tag{12}$$

each cross derivative in (iii) decomposes into a *collider contribution* T_C from the propensity term and a *bidirected contribution* T_B from the marginal Z-U dependence (see Fig. 2). Under our assumptions, these two contributions carry opposite signs, so each inequality reduces to a question of which dominates. Consider first the within-U part of Eq. 11. Under monotonicity of π in u (assumption (i)), T_C tends to be log-submodular in u, contributing a non-positive cross derivative. By assumption (ii), T_B is log-supermodular in u, contributing a non-negative one. The within-U inequality therefore asks that T_B dominates: the positive marginal dependence among the U coordinates must be strong enough to survive the collider's re-weighting through T_C. The across-Z-U inequality flips both the required sign and the dominant channel. Under (i), the cross derivative of T_C in z-u is typically non-positive (the collider couples Z and U negatively at fixed X); whenever Z and U are marginally positively dependent, the cross derivative of T_B is non-negative. The across-Z-U inequality therefore asks that T_C dominates: conditioning on $\{X = x\}$ must be strong enough to flip the marginal positive Z-U dependence into a net negative one.

In the context of our ICU example, condition (iii) constrains how the propensity for mechanical ventilation interacts with the marginal dependence between observed severity (e.g., SOFA score, oxygen saturation) and unobserved severity (e.g., perceived difficulty breathing, hemodynamic instability not captured in Z). It does not require Z and U to be independent, and the two-channel decomposition above gives a concrete reading of when it is more or less plausible. A natural regime in which (iii) is satisfied is one where Z and U correspond to different organ systems, e.g., Z captures cardiovascular markers while U captures respiratory markers. In such a setting, intra-system severities (within U, e.g., difficulty breathing and respiratory rate) tend to co-vary tightly, supporting the within-U inequality, while cross-system severities (between Z and U) are positively but more loosely related, leaving room for the collider channel to overturn the marginal Z-U dependence. The opposite regime is also possible: if Z and U relate to the same organ system, the marginal Z-U dependence may be strong enough that the collider cannot overturn it, in which case the across-Z-U inequality fails and Z-dominance offers no guarantee. Whether (iii) holds is therefore an empirical question about how the observed and unobserved severity markers partition across systems, and we view it as an assumption the analyst should reason about in light of the specific covariates available, rather than one that is guaranteed by the clinical setting alone. Put formally, Prop. 4 may fail if there is either strong correlation of Z, U or a strong interaction of Z, U in f_X (see Exs. 5 and 6). Finally, we remark that the chain argument of Prop. 4 extends to π-matching with a single, scalar version of assumption (iii), as discussed in App. C.

2.4 Extraction Strategies

We now turn to the second part of our framework: how the proposed pair (i, j) is processed by the expert annotator. We decompose this into three stages: extraction, selection, and success. Together, these stages determine the probability that the elicited explanation corresponds to a genuine UC.

Extraction. Given a pair (i, j) with $X_i = 1, X_j = 0$, the expert produces a candidate set $C_\varepsilon(i, j)$ of explanations for the differential treatment. We emphasize that an explanation is inherently *contrastive*: a variable $V^{(l)}$ (observed or unobserved) is a candidate explanation only if $V_i^{(l)} > V_j^{(l)}$ and $V^{(l)}$ influences X. The set $C_\varepsilon(i, j)$ is therefore a subset of the variables along which i exceeds j, generated according to an *extraction strategy* ε. Under *perfect extraction* $\varepsilon_{\text{perf}}$, the extracted set correctly includes all such variables: $C_{\varepsilon_{\text{perf}}}(i, j) = \{V^{(l)} : V_i^{(l)} > V_j^{(l)}\}$. In a more general setting, such as *ablation-based extraction* ε_{abl} (natural when the expert is a large language model performing counterfactual reasoning over textual notes), $C_\varepsilon(i, j)$ may contain *hallucinated* variables that did not drive the decision, or *omit* genuine drivers that influenced the decision.

Selection. Given $C_\varepsilon(i,j)$, we assume the expert selects a single explanation uniformly at random among all candidates they consider plausible. These include unobserved candidates $U^{(l)}$, and observed coordinates $Z^{(k)}$ along which i exceeds j. The probability that the expert selects an unobserved variable ($E=1$) is therefore

$$P(E=1 \mid i,j) = \frac{|C_\varepsilon(i,j) \setminus Z|}{|C_\varepsilon(i,j)|}. \tag{13}$$

Success. Selection alone is not sufficient, as the selected variable must correspond to a genuine UC to be informative. We say the extraction is *accurate* ($A=1$) when $E=1$ and the selected variable c satisfies $U_i^{(c)} > U_j^{(c)}$. Under perfect extraction, accuracy simplifies to

$$P(A=1 \mid i,j,\varepsilon_{\text{perf}}) = \frac{N(i,j)}{N(i,j)+D(i,j)}, \tag{14}$$

with $N(i,j) = |\{U^{(l)} : U_i^{(l)} > U_j^{(l)}\}|$ counting the unobserved drivers, and $D(i,j) = |\{Z^{(k)} : Z_i^{(k)} > Z_j^{(k)}\}|$ counting the competing observed explanations. A good matching strategy should maximize N (number of active UCs) and minimize D (number of active known confounders). Our first result under the rubric of *strategy dominance* shows that Z-dominance provably increases accuracy compared to Z-matching under perfect extraction:

Theorem 2 (Z-dominance dominates Z-matching)**.** *Fix $Z=z$ and assume conditions (i)–(iii) of Prop. 4 hold at $X=1$. Then*

$$\mathbb{E}[A \mid \mathcal{M}^{Z\text{-}dom}, Z_i = z', Z_j = z, \varepsilon_{\text{perf}}] \;\geq\; \mathbb{E}[A \mid \mathcal{M}^{Z\text{-}match}, Z_i = Z_j = z, \varepsilon_{\text{perf}}] \;\forall z' < z. \tag{15}$$

2.5 π-Matching Beats Marginal Matching

While Thm. 2 establishes strategy dominance for Z-dominance over Z-matching, proving a similar result for π-matching via the non-linear success probability $N/(N+D)$ requires strong assumptions over a two-dimensional $(N,-D)$ distribution. To make strategy comparisons more tractable, we introduce a linearized surrogate utility

$$\mathcal{U}_\alpha(\mathcal{M}) \;\coloneqq\; \mathbb{E}[N - \alpha D \mid \mathcal{M}], \qquad \alpha > 0, \tag{16}$$

which retains a qualitative interpretation similar to $N/(N+D)$. For our analysis, we quantify a variable's predictive power via its marginal Area Under Curve (AUC), defined inline as $\mathrm{AUC}(V^{(k)}) \coloneqq P(V_i^{(k)} > V_j^{(k)} \mid X_i=1, X_j=0)$ for independent draws $V_i \sim P(V \mid X=1)$ and $V_j \sim P(V \mid X=0)$. For the observables, $\mathrm{AUC}(Z^{(k)})$ values can be computed, while for the UCs, $\mathrm{AUC}(U^{(l)})$ are unknown. AUC values equal to $1/2$ correspond to independence $V^{(k)} \perp\!\!\!\perp X$.

Theorem 3 (π-matching Dominates Marginal Matching under $\mathcal{U}_\alpha$)**.** *Assume that the conditions of Prop. 3 hold and that each marginal $Z^{(k)}$ is continuous. Then*

$$\mathcal{U}_\alpha(\mathcal{M}^{\pi\text{-}match}) - \mathcal{U}_\alpha(\mathcal{M}^{\text{marg}}) \;\geq\; \alpha \sum_{k=1}^{d_Z} \big(\mathrm{AUC}(Z^{(k)}) - \tfrac{1}{2}\big) - \sum_{l=1}^{d_U} \big(\mathrm{AUC}(U^{(l)}) - \tfrac{1}{2}\big). \tag{17}$$

This theorem provides actionable insight into when π-matching outperforms marginal matching. Under Prop. 3, π-matching guarantees a lower bound of $\mathbb{E}[N \mid \mathcal{M}^{\pi\text{-match}}] \geq d_U/2$. While this expectation is generally smaller than marginal matching's $\mathbb{E}[N \mid \mathcal{M}^{\text{marg}}] = \sum_{l=1}^{d_U} \mathrm{AUC}(U^{(l)})$, π-matching reduces the expected number of observed competitors strictly to $\mathbb{E}[D \mid \mathcal{M}^{\pi\text{-match}}] = d_Z/2$, compared to $\mathbb{E}[D \mid \mathcal{M}^{\text{marg}}] = \sum_{k=1}^{d_Z} \mathrm{AUC}(Z^{(k)})$. Thm. 3 gives us that

$$\frac{\sum_{l=1}^{d_U} \big(\mathrm{AUC}(U^{(l)}) - \frac{1}{2}\big)}{\sum_{k=1}^{d_Z} \big(\mathrm{AUC}(Z^{(k)}) - \frac{1}{2}\big)} \leq 1 \implies \mathbb{E}[N-D \mid \mathcal{M}^{\pi\text{-match}}] \geq \mathbb{E}[N-D \mid \mathcal{M}^{\text{marg}}]. \tag{18}$$

In other words, π-matching dominates when the amount of variation explained by Z is large compared to the variation explained by U (e.g., when $d_Z \gg d_U$, meaning most confounders are known; or when $Z^{(k)}$ AUCs are large compared to $U^{(l)}$). We verify this in the experiments.

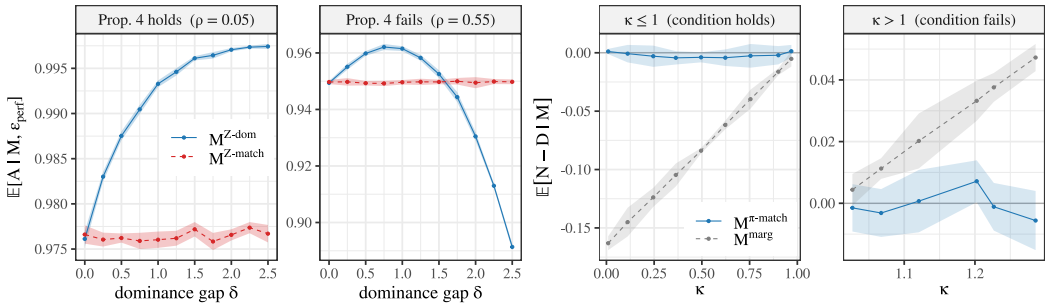


Figure 3: Synthetic verification. *Panels 1–2*: $\mathcal{M}^{Z\text{-dom}}$ vs. $\mathcal{M}^{Z\text{-match}}$ as a function of the dominance gap δ, with (iii) holding ($\rho = 0.05$) and failing ($\rho = 0.55$). *Panels 3–4*: $\mathcal{M}^{\pi\text{-match}}$ vs. $\mathcal{M}^{\text{marg}}$ as a function of κ, in the regimes $\kappa \leq 1$ and $\kappa > 1$. Bands are 95% across-rep CIs.

3 Experiments

We verify Thms. 2 and 3 on synthetic data (Sec. 3.1), and evaluate our framework on a semi-synthetic dataset derived from MIMIC-III (Sec. 3.2), while App. F analyzes the real MIMIC-III data.

3.1 Synthetic verification

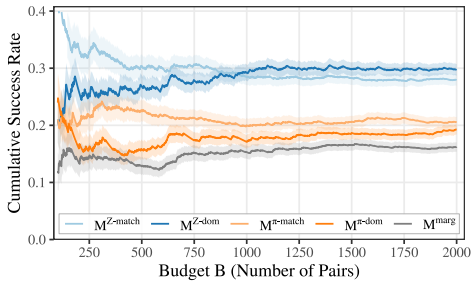
In both experiments, (Z, U) are Gaussians $\mathcal{N}(0, \Sigma)$ with $d_Z = d_U = 3$, Z-block has $\Sigma_Z = I$, U-block $\Sigma_U = \sigma_\varepsilon^2 I + \tau^2 \mathbf{1}\mathbf{1}^\top$. Setting $\sigma_\varepsilon^2 = 0.4$, and $\tau^2 = 0.6$ makes $P(U \mid Z = z)$ log-supermodular, and propensity is given by $\pi(z, u) = \sigma(\alpha + \beta^\top z + \gamma^\top u)$ with $\alpha = -1$. The cross-block covariance Σ_{ZU} is the key parameter determining whether assumptions of Props. 3 and 4 hold true. We work under $\varepsilon_{\text{perf}}$ throughout and estimate expectations by Monte Carlo (Fig. 3).

Thm. 2 (Panels 1–2). With $\beta = \gamma = \mathbf{1}$, $z = 0$, and $z' = -\delta \cdot \mathbf{1}$, we estimate $\mathbb{E}[A \mid \mathcal{M}, \varepsilon_{\text{perf}}]$ for $\mathcal{M}^{Z\text{-match}}$ and $\mathcal{M}^{Z\text{-dom}}$ by sampling $U_i \sim P(U \mid Z, X = 1)$ and $U_j \sim P(U \mid Z = z, X = 0)$ via rejection from $\mathcal{N}(\mu(z), \Sigma_{U|Z})$ with acceptance $\pi^X (1 - \pi)^{1-X}$. Under $\varepsilon_{\text{perf}}$, both strategies have $D = 0$, so $A = \mathbb{1}\{N \geq 1\}$. In Panel 1, $\Sigma_{ZU} = \rho I$ with $\rho = 0.05$, which means mild Z-U dependence, and that (iii) holds at $X = 1$. In Panel 2, $\rho = 0.55$, implying strong Z-U dependence, and the across-Z-U inequality flips at $X = 1$. Panel 1 confirms Eq. 15: $\mathcal{M}^{Z\text{-dom}}$ dominates $\mathcal{M}^{Z\text{-match}}$ as δ grows under Prop. 4. Panel 2 illustrates the failure mode: when condition (iii) is violated, for larger values of the dominance gap δ, Z-matching outperforms Z-dominance.

Thm. 3 (Panels 3–4). Using the same parametric family, we fix $\beta = 0.2 \cdot \mathbf{1}$, $\gamma = 0$ and sweep $\Sigma_{ZU} = c \cdot \mathbf{1}\mathbf{1}^\top$ with $c \in [0, 0.43]$ (upper end keeps $P(U \mid Z)$ log-supermodular). Each c pins down a value of the AUC ratio $\kappa := \sum_l (\text{AUC}(U^{(l)}) - \frac{1}{2}) / \sum_k (\text{AUC}(Z^{(k)}) - \frac{1}{2})$ appearing in Eq. 18, which we estimate from a simulated population ($n = 10^5$) using random treated/untreated pairs. We then collect $\mathcal{M}^{\pi\text{-match}}$ pairs by rejection on $|\pi(Z_i) - \pi(Z_j)| < 0.5\%$, with $\pi(Z)$ computed by marginalizing out U in the σ function. We plot $\mathbb{E}[N - D \mid \mathcal{M}]$ for both $\mathcal{M}^{\pi\text{-match}}$ and $\mathcal{M}^{\text{marg}}$ as a function of κ. Parameter $\gamma = 0$ is chosen so Eq. 17 holds with equality and the threshold $\kappa = 1$ is sharp, meaning that $\mathbb{E}[N - D \mid \mathcal{M}^{\pi\text{-match}}] - \mathbb{E}[N - D \mid \mathcal{M}^{\text{marg}}] = (1 - \kappa) \sum_k (\text{AUC}(Z^{(k)}) - \frac{1}{2})$, so the linearized utility gap $\mathcal{U}_\alpha(\mathcal{M}^{\pi\text{-match}}) - \mathcal{U}_\alpha(\mathcal{M}^{\text{marg}})$ from Eq. 16 changes sign exactly at $\kappa = 1$. Panel 3 for $\kappa \leq 1$ confirms this, with $\mathcal{M}^{\pi\text{-match}}$ above $\mathcal{M}^{\text{marg}}$, as predicted. Panel 4 for $\kappa > 1$ shows that the inequality indeed reverses, as predicted by the theoretical result.

3.2 Semi-synthetic experiments on MIMIC-III

Data construction. We build MIMIC-III-SeS by using 12 observed covariates Z (resp. rate, MAP, lactate, P/F ratio, pCO_2, pO_2, O_2 saturation, age, sex, Charlson, SOFA, plus admission-diagnosis indicators). The unobserved confounders U have dimension $d_U = 10$, and all are binary, extracted from clinical notes via UMLS entity linking with negation detection (Neumann et al., 2019). The concepts include pleural effusion, heart failure, dyspnea, pneumonia, pulmonary edema, bowel



Adjustment	Estimated ETT mean $\pm 2\widehat{\mathrm{SE}}$	Signed ETT bias mean $\pm 2\widehat{\mathrm{SE}}$
$\mathcal{M}^{\pi\text{-match}}$ top-10	$-1.5\% \pm 0.11\%$	$+0.2\% \pm 0.11\%$
$\mathcal{M}^{\mathrm{marg}}$ top-10	$-0.9\% \pm 0.14\%$	$+0.8\% \pm 0.14\%$
Z-only	$+1.1\% \pm 0.10\%$	$+2.8\% \pm 0.10\%$
Oracle: $Z + U$	$-1.8\% \pm 0.11\%$	$0.0\% \pm 0.11\%$

Figure 4: Results on MIMIC-III-SeS. *Left*: cumulative success rate over increasing comparison budget B. *Right*: downstream ETT recovery after measuring the top-10 prioritized confounders.

sound presence, hypoxia, atrial fibrillation, hypertensive disease, and atelectasis. We also add unobserved false-positive variables U^{fp}, constructed by independently permuting values of each $U^{(l)}$ across patients. We fit a logistic propensity $\pi(Z, U) = \sigma(\alpha_X + \beta_X^\top Z + \gamma_X^\top U + \gamma_{\mathrm{fp},X}^\top U^{\mathrm{fp}})$ with $\gamma_X^{(l)} \sim \mathrm{Unif}[0.3, 0.4]$ (so π is monotone in u) and $\gamma_{\mathrm{fp},X} = 0$ (false positives have no effect on X). The outcome follows a logistic model with coefficients $\gamma_Y^{(l)} \sim \mathrm{Unif}[0.3, 0.4]$ for U and $\gamma_{X\to Y} = -0.2$ for X, implying treatment X is protective. We sample new $(\tilde{X}, \tilde{Y})$ from this model, making data $(Z, \tilde{X}, \tilde{Y})$ semi-synthetic. U and U^{fp} are never exposed to the matching strategies.

Strategy implementation & performance. For a practical implementation of $\mathcal{M}^{Z\text{-match}}$, we compute the Euclidean distance on standardized Z values for every pair (i, j) with $X_i = 1, X_j = 0$, sort in ascending order, and take the top B pairs, selecting pairs closest in Z-space. $\mathcal{M}^{Z\text{-dom}}$ counts the coordinates where i exceeds j (margin 0.2σ), and picks pairs where this score is smallest. $\mathcal{M}^{\pi\text{-match}}$ considers ascending $|\hat{\pi}(Z_i) - \hat{\pi}(Z_j)|$ values, with $\hat{\pi}$ estimated by 5-fold cross-validated logistic regression on Z; $\mathcal{M}^{\pi\text{-dom}}$ (see App. C) considers ascending $\hat{\pi}(Z_i) - \hat{\pi}(Z_j)$. $\mathcal{M}^{\mathrm{marg}}$ considers a random order over the data. We vary the budget B, which represents the number of pairs proposed by a strategy. For each selected pair we compute the probability of success $\lambda(i, j) \coloneqq P(A = 1 \mid i, j, \varepsilon_{\mathrm{perf}})$ as in Eq. 14. The cumulative average of λ over B, shown in Fig. 4 (left panel), traces how successful the strategies are with increasing budget sizes. The Z-based strategies $\mathcal{M}^{Z\text{-match}}$ and $\mathcal{M}^{Z\text{-dom}}$ achieve the highest success, followed by $\mathcal{M}^{\pi\text{-match}}$ and $\mathcal{M}^{\pi\text{-dom}}$, while $\mathcal{M}^{\mathrm{marg}}$ has lowest success throughout. These results empirically verify Thms. 2 and 3.

Downstream ETT recovery. For each matching strategy we define the hit score $H_l^M(B) \coloneqq \sum_{b=1}^{B} \mathbb{1}\{U_{i_b}^{(l)} > U_{j_b}^{(l)}\}$ corresponding to the number of times variable l is extracted based on matched pairs (i_b, j_b). Let $U_{\text{Top-10}}^M(B)$ be the 10 variables with the highest hit rates $H_l^M(B)$ after budget B. For $\mathcal{M}^{\mathrm{marg}}$ and $\mathcal{M}^{\pi\text{-match}}$, we compare the ETT obtained by adjusting for $Z, U_{\text{Top-10}}^M$, against adjusting for Z-only and the oracle adjustment for Z, U. The results are shown in Fig. 4 (right panel), indicating $\mathcal{M}^{\pi\text{-match}}$ performs best and substantially reduces UC bias ($\mathcal{M}^{\mathrm{marg}}$ also reduces UC bias, although not to the same degree), showing effect recovery is possible even when the (Z, U) structure is inherited from real-world clinical data. Further analyses of MIMIC-III-SeS are in App. E.

Limitations and future work. Several aspects of our framework merit further development. First, the multivariate stochastic-dominance notion underlying Thm. 1 is strong; relaxing this notion to coordinate-wise marginal stochastic dominance is a natural next step that would likely allow relaxing the required assumptions (such as log-supermodularity of $P(u \mid z)$). Second, our analysis of extraction (Sec. 2.4) rests on a specific model of expert behavior with uniform selection (Eq. 14); this is a first formalization, and we expect real-world elicitation may work differently. Third, the framework presupposes that the decision-maker can articulate the factors driving their decisions when comparing two units; in settings where treatment is allocated on tacit or subconscious cues, the elicited explanations may underrepresent the true confounders. Finally, our framework detects *candidates* for unobserved confounders with a probabilistic guarantee, but does not certify whether any specific detected variable is a confounder of $X \to Y$, nor does it establish that adding the detected variables recovers back-door validity. We leave these interesting directions for future work.

AI USE STATEMENT

We used AI tools to correct grammar and typographical errors and to assist with plotting. All other aspects of the work, including the research methodology, main implementation, experiments, and interpretation of results, were carried out by the authors. We reviewed and verified all AI-assisted edits, plotting code, and resulting figures for correctness and consistency with the underlying results.

ETHICS STATEMENT

This work develops a methodology for identifying candidate unobserved confounders in observational studies, evaluated using retrospective ICU data and synthetic or semi-synthetic experiments. It does not involve interventions in patient care. Clinical data require careful protection of patient privacy, particularly when using information extracted from clinical notes. Our findings should not be interpreted as clinical treatment recommendations: apparent harmful treatment effects in observational data may reflect unobserved confounding. Furthermore, elicited explanations may reflect incomplete knowledge or biases of the decision-maker, and identifying candidate confounders does not guarantee valid causal conclusions. Any future clinical application would require independent validation, appropriate data governance, and assessment of performance across patient groups. The intended contribution is to improve the reliability of observational research by making potential sources of confounding more explicit.

REPRODUCIBILITY STATEMENT

The source code for reproducing our results is available in the anonymized repository at `https://anonymous.4open.science/r/icu-deconfounding-5DB9`, together with a README describing the environment setup and instructions for running the experiments. The main paper specifies the theoretical assumptions and experimental procedures, while the supplementary material provides proofs and additional experimental details.

REFERENCES

Susan Athey and Stefan Wager. Estimating treatment effects with causal forests: An application. *Observational studies*, 5(2):37–51, 2019.

Susan Athey, Guido W Imbens, et al. Combining experimental and observational data to estimate treatment effects on long-term outcomes. *arXiv preprint arXiv:2006.09676*, 2020.

Elias Bareinboim. *Causal Artificial Intelligence: A Roadmap for Building Causally Intelligent Systems*. Online, 2025. URL `https://causalai-book.net/`. Draft version.

Elias Bareinboim, Andrew Forney, and Judea Pearl. Bandits with unobserved confounders: A causal approach. *Advances in neural information processing systems*, 28, 2015.

Elias Bareinboim, Juan D. Correa, Duligur Ibeling, and Thomas Icard. On pearls hierarchy and the foundations of causal inference. In *Probabilistic and Causal Inference: The Works of Judea Pearl*, pp. 507556. Association for Computing Machinery, New York, NY, USA, 1st edition, 2022.

Roger C Bone, Robert A Balk, Frank B Cerra, R Phillip Dellinger, Alan M Fein, William A Knaus, Roland MH Schein, and William J Sibbald. Definitions for sepsis and organ failure and guidelines for the use of innovative therapies in sepsis. *Chest*, 101(6):1644–1655, 1992.

Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pp. 785–794, 2016.

Andres Esteban, Antonio Anzueto, Inmaculada Alia, Federico Gordo, Carlos Apezteguia, Fernando Palizas, David Cide, Rosanne Goldwaser, Luis Soto, Guillermo Bugedo, et al. How is mechanical ventilation employed in the intensive care unit? an international utilization review. *American journal of respiratory and critical care medicine*, 161(5):1450–1458, 2000.

Michael O Harhay, Jason Wagner, Sarah J Ratcliffe, Rachel S Bronheim, Anand Gopal, Sydney Green, Elizabeth Cooney, Mark E Mikkelsen, Meeta Prasad Kerlin, Dylan S Small, et al. Outcomes and statistical power in adult critical care randomized trials. *American journal of respiratory and critical care medicine*, 189(12):1469–1478, 2014.

William R Hersh, Mark G Weiner, Peter J Embi, Judith R Logan, Thomas H Payne, Elmer V Bernstam, Harold P Lehmann, George Hripcsak, Timothy H Hartzog, James J Cimino, et al. Caveats for the use of operational electronic health record data in comparative effectiveness research. *Medical care*, pp. S30–S37, 2013.

Ashish K Jha, Catherine M DesRoches, Eric G Campbell, Karen Donelan, Sowmya R Rao, Timothy G Ferris, Alexandra Shields, Sara Rosenbaum, and David Blumenthal. Use of electronic health records in us hospitals. *New England Journal of Medicine*, 360(16):1628–1638, 2009.

Alistair EW Johnson, Tom J Pollard, Lu Shen, Li-Wei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3:160035, 2016. doi: 10.1038/sdata.2016.35. URL https://www.nature.com/articles/sdata201635.

Samuel Karlin and Yosef Rinott. Classes of orderings of measures and related correlation inequalities. i. multivariate totally positive distributions. *Journal of Multivariate Analysis*, 10(4):467–498, 1980.

Sanghack Lee, Juan D Correa, and Elias Bareinboim. General identifiability with arbitrary surrogate experiments. In *Uncertainty in artificial intelligence*, pp. 389–398. PMLR, 2020.

Mark Neumann, Daniel King, Iz Beltagy, and Waleed Ammar. Scispacy: fast and robust models for biomedical natural language processing. In *Proceedings of the 18th BioNLP workshop and shared task*, pp. 319–327, 2019.

Judea Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, New York, 2000. 2nd edition, 2009.

Nils Rodemund, Bernhard Wernly, Christian Jung, et al. The salzburg intensive care database (sicdb): an openly available critical care dataset. *Intensive Care Medicine*, 49:700–702, 2023. doi: 10.1007/s00134-023-07046-3. URL https://link.springer.com/article/10.1007/s00134-023-07046-3.

Paul R Rosenbaum and Jeffrey H Silber. Matching and thick description in an observational study of mortality after surgery. *Biostatistics*, 2(2):217–232, 2001.

Evan TR Rosenman, Guillaume Basse, Art B Owen, and Mike Baiocchi. Combining observational and experimental datasets using shrinkage estimators. *Biometrics*, 79(4):2961–2973, 2023.

Kenneth F Schulz. Randomised trials, human nature, and reporting guidelines. *The Lancet*, 348 (9027):596–598, 1996.

Peter J Thoral, Johannes M Peppink, Rob H Driessen, et al. Sharing icu patient data responsibly under the society of critical care medicine/european society of intensive care medicine joint data science collaboration: the amsterdam university medical centers database (amsterdamumcdb) example. *Critical Care Medicine*, 2021. doi: 10.1097/CCM.0000000000004916. URL https://journals.lww.com/ccmjournal/fulltext/2021/06000/Sharing_ICU_Patient_Data_Responsibly_Under_the.16.aspx?context=FeaturedArticles&collectionId=2.

SUPPLEMENTARY MATERIAL: *Confounder Detection via Treatment Intent: A New Observational Study Design*

The source code for reproducing the results can be found in the anonymized code repository `https://anonymous.4open.science/r/icu-deconfounding-5DB9`. The repository includes a README file explaining the setup steps. Most experiments were run on a MacBook Pro M3 (Tahoe 26.2) with under 24 hours of total computation.

A PROOFS

In this appendix, we provide formal proofs of the results appearing in the main text.

Proof. Fix z and let $\pi(z,u) = P(X=1 \mid Z=z, U=u)$. By Bayes' rule,

$$P(u \mid Z=z, X=1) \propto \pi(z,u)\, P(u \mid Z=z), \tag{19}$$

$$P(u \mid Z=z, X=0) \propto (1-\pi(z,u))\, P(u \mid Z=z). \tag{20}$$

The likelihood ratio is therefore

$$r(u) \coloneqq \frac{P(u \mid Z=z, X=1)}{P(u \mid Z=z, X=0)} \propto \frac{\pi(z,u)}{1-\pi(z,u)}, \tag{21}$$

which is non-decreasing in u since $\pi(z,u)$ is. Hence the family $\{P(\cdot \mid Z=z, X=x)\}_{x\in\{0,1\}}$ has monotone likelihood ratio in u, which implies the stochastic ordering in Eq. 7. □

We first state a key lemma of Karlin and Rinott (Karlin & Rinott, 1980) required for the proof of Prop. 2:

Lemma 4 (Karlin-Rinott, 1980 (Karlin & Rinott, 1980))**.** *Let f_0, f_1 be densities on $\mathbb{R}^d$ satisfying*

$$f_0(u)\, f_1(u') \;\le\; f_0(u \wedge u')\, f_1(u \vee u') \quad \textit{for all } u, u' \in \mathbb{R}^d, \tag{22}$$

Then $f_1 \succeq_{st} f_0$.

Proof of Prop. 2. Fix z and write $\pi(u) = P(X=1 \mid Z=z, U=u)$, $g(u) = P(u \mid Z=z)$. By Bayes' rule,

$$P(u \mid Z=z, X=1) \propto \pi(u)\, g(u), \qquad P(u \mid Z=z, X=0) \propto (1-\pi(u))\, g(u). \tag{23}$$

We verify the Karlin-Rinott condition of Lem. 4. By assumption (i), $\pi(u \vee u') \ge \pi(u')$ and $1 - \pi(u \wedge u') \ge 1 - \pi(u)$, so

$$(1-\pi(u))\, \pi(u') \;\le\; (1-\pi(u \wedge u'))\, \pi(u \vee u'). \tag{24}$$

Combining Eq. 24 with assumption (ii),

$$(1-\pi(u))\, g(u) \cdot \pi(u')\, g(u') \;\le\; (1-\pi(u \wedge u'))\, g(u \wedge u') \cdot \pi(u \vee u')\, g(u \vee u'), \tag{25}$$

which is exactly the condition of Lem. 4 applied to $f_0(u) \propto (1-\pi(u))g(u)$ and $f_1(u) \propto \pi(u)g(u)$, and the conclusion follows. □

Proof of Prop. 3. By the covariate balancing property of the propensity score, $X \perp\!\!\!\perp Z \mid \pi(Z)$, which implies $P(Z \mid \pi(Z)=p, X=x) = P(Z \mid \pi(Z)=p)$ for $x \in \{0,1\}$. Expanding over Z gives:

$$P(U \mid \pi(Z)=p, X=x) = \int P(U \mid Z=z, X=x)\, P(z \mid \pi(Z)=p)\, dz. \tag{26}$$

Both treatment groups are mixtures with respect to the same mixing distribution $P(z \mid \pi(Z)=p)$. By Prop. 2, $P(U \mid Z=z, X=1) \succeq_{st} P(U \mid Z=z, X=0)$ for every z. Stochastic dominance

is preserved under mixing with a common mixing measure, since for every coordinatewise non-decreasing ϕ,

$$\mathbb{E}[\phi(U) \mid \pi(Z) = p, X = 1] = \int \mathbb{E}[\phi(U) \mid Z = z, X = 1]\, P(z \mid \pi(Z) = p)\, dz \tag{27}$$

$$\geq \int \mathbb{E}[\phi(U) \mid Z = z, X = 0]\, P(z \mid \pi(Z) = p)\, dz \tag{28}$$

$$= \mathbb{E}[\phi(U) \mid \pi(Z) = p, X = 0], \tag{29}$$

which implies Eq. 9. □

Proof of Prop. 4. The target claim is decomposed into two-step chains:

$$P(U \mid z', X{=}1) \ \succeq_{st} \ P(U \mid z', X{=}0) \ \succeq_{st} \ P(U \mid z, X{=}0), \tag{30}$$

$$P(U \mid z', X{=}1) \ \succeq_{st} \ P(U \mid z, X{=}1) \ \succeq_{st} \ P(U \mid z, X{=}0). \tag{31}$$

In each chain, one link follows from Prop. 2 under (i) and (ii) (applied at $Z = z'$ in Eq. 30, at $Z = z$ in Eq. 31). The other link is the Z-shift Eq. 10, which we now establish via assumption. We treat the $x = 0$ case; $x = 1$ is symmetric with π in place of $1 - \pi$.

Set $f_0(u) \coloneqq P(u \mid z, X = 0)$ and $f_1(u) \coloneqq P(u \mid z', X = 0)$. Both are proportional to $h^{(0)}(\cdot, u)$ at the corresponding z argument, with u-independent normalizers. We want to show that, for all $u, u' \in \mathbb{R}^d$,

$$f_0(u)\, f_1(u') \ \leq \ f_0(u \wedge u')\, f_1(u \vee u'); \tag{32}$$

equivalently, after canceling normalizers,

$$h^{(0)}(z, u)\, h^{(0)}(z', u') \ \leq \ h^{(0)}(z, u \wedge u')\, h^{(0)}(z', u \vee u'). \tag{33}$$

For $u \leq u'$ or $u \geq u'$ coordinatewise, Eq. 33 is trivial. For non-comparable u, u', the derivative conditions in Eq. 11 at $X = 0$ imply, by integration along an axis-aligned path from $(z', u \vee u', z, u \wedge u')$ to (z, u, z', u'), the four-point inequality

$$\log h^{(0)}(z, u) + \log h^{(0)}(z', u') \ \leq \ \log h^{(0)}(z, u \wedge u') + \log h^{(0)}(z', u \vee u'), \tag{34}$$

which is exactly Eq. 33. Applying Lem. 4 to Eq. 32 yields $f_1 \succeq_{st} f_0$, which is the $X = 0$ link of Eq. 30, completing the chain. □

Proof of Thm. 2. Note that for both $\mathcal{M}^{Z\text{-match}}$ and $\mathcal{M}^{Z\text{-dom}}$, we have $|\{Z^{(k)} : Z_i^{(k)} > Z_j^{(k)}\}| = 0$, i.e., $D(i, j) = 0$. Therefore, $A = \mathbf{1}\{N \geq 1\}$ for both strategies, and the comparison reduces to $P(N \geq 1)$ (a single unobserved confounder greater for i than j is sufficient for success). Conditional on (Z_i, Z_j, X_i, X_j), the variables U_i and U_j are independent with $U_i \sim P(\cdot \mid Z_i, X_i = 1)$ and $U_j \sim P(\cdot \mid Z_j, X_j = 0)$. Define

$$\psi(u; U_j) \coloneqq \mathbf{1}\{\exists\, l : u^{(l)} > U_j^{(l)}\}, \tag{35}$$

which is a coordinatewise non-decreasing function in u for every fixed value of U_j. Hence $u \mapsto \mathbb{E}_{U_j}[\psi(u; U_j)]$ is also coordinatewise non-decreasing. For the untreated unit, both strategies condition on $Z_j = z, X_j = 0$ and thus $P(U_j \mid Z_j = z, X_j = 0)$ is the same between the strategies. However, the distribution of U_i changes from $P(\cdot \mid z, X = 1)$ under $\mathcal{M}^{Z\text{-match}}$ to $P(\cdot \mid z', X = 1)$ under $\mathcal{M}^{Z\text{-dom}}$. By the $X = 1$ chain link (Eq. 31) of Prop. 4, which uses condition (iii) at $X = 1$,

$$P(U \mid Z = z', X = 1) \ \succeq_{st} \ P(U \mid Z = z, X = 1). \tag{36}$$

Multivariate stochastic dominance applied to the coordinatewise non-decreasing function $u \mapsto \mathbb{E}_{U_j}[\psi(u; U_j)]$, together with noting that $\mathbb{E}[\psi \mid C] = P(N \geq 1 \mid C)$ conditional on any C, gives the claim in Eq. 15. □

Proof of Thm. 3. Decompose into N and D contributions:

$$\mathcal{U}_\alpha(\mathcal{M}^{\pi\text{-match}}) - \mathcal{U}_\alpha(\mathcal{M}^{\text{marg}}) \ = \ \underbrace{\big(\mathbb{E}[N \mid \mathcal{M}^{\pi\text{-match}}] - \mathbb{E}[N \mid \mathcal{M}^{\text{marg}}]\big)}_{=:\,\Delta_N} \tag{37}$$

$$- \ \alpha \underbrace{\big(\mathbb{E}[D \mid \mathcal{M}^{\pi\text{-match}}] - \mathbb{E}[D \mid \mathcal{M}^{\text{marg}}]\big)}_{=:\,\Delta_D}. \tag{38}$$

By linearity, $\mathbb{E}[N \mid \mathcal{M}] = \sum_l P(U_i^{(l)} > U_j^{(l)} \mid \mathcal{M})$ and $\mathbb{E}[D \mid \mathcal{M}] = \sum_k P(Z_i^{(k)} > Z_j^{(k)} \mid \mathcal{M})$.

Marginal matching. Under $\mathcal{M}^{\text{marg}}$, $(Z_i, U_i) \sim P(\cdot \mid X = 1)$ and $(Z_j, U_j) \sim P(\cdot \mid X = 0)$ are independent. The per-coordinate probabilities therefore equal the corresponding marginal selection AUCs by definition:

$$\mathbb{E}[D \mid \mathcal{M}^{\text{marg}}] = \sum_{k=1}^{d_Z} \text{AUC}(Z^{(k)}), \qquad \mathbb{E}[N \mid \mathcal{M}^{\text{marg}}] = \sum_{l=1}^{d_U} \text{AUC}(U^{(l)}). \tag{39}$$

π-match, D-side. Conditional on $\pi(Z_i) = \pi(Z_j) = p$, the covariate-balancing property $X \perp\!\!\!\perp Z \mid \pi(Z)$ implies Z_i and Z_j are i.i.d. draws from $P(Z \mid \pi(Z) = p)$. Continuity of the marginal of $Z^{(k)}$ and exchangeability give

$$P(Z_i^{(k)} > Z_j^{(k)} \mid \pi = p) \;=\; \tfrac{1}{2} \tag{40}$$

for every k and p. Summing over k and averaging over the pair distribution of π yields $\mathbb{E}[D \mid \mathcal{M}^{\pi\text{-match}}] = d_Z/2$. Hence

$$\Delta_D \;=\; \tfrac{d_Z}{2} - \sum_k \text{AUC}(Z^{(k)}) \;=\; -\sum_k \big(\text{AUC}(Z^{(k)}) - \tfrac{1}{2}\big). \tag{41}$$

π-match, N-side. Conditional on $\pi = p$, Prop. 3 gives the multivariate ordering $P(U \mid \pi = p, X = 1) \succeq_{st} P(U \mid \pi = p, X = 0)$, which implies the marginal ordering on each coordinate:

$$P(U^{(l)} \mid \pi = p, X = 1) \;\succeq_{st}\; P(U^{(l)} \mid \pi = p, X = 0). \tag{42}$$

For two independent univariate continuous random variables A, B with $A \succeq_{st} B$ and CDFs $F_A \leq F_B$,

$$P(A > B) \;=\; \mathbb{E}[1 - F_A(B)] \;\geq\; \mathbb{E}[1 - F_B(B)] \;=\; \tfrac{1}{2}, \tag{43}$$

where the last equality uses the probability integral transform $F_B(B) \sim \text{Unif}[0, 1]$. Applied to Eq. 42 conditionally on $\pi = p$, this yields $P(U_i^{(l)} > U_j^{(l)} \mid \pi = p) \geq 1/2$. Summing over l and averaging over p,

$$\mathbb{E}[N \mid \mathcal{M}^{\pi\text{-match}}] \;\geq\; d_U/2, \tag{44}$$

and therefore

$$\Delta_N \;\geq\; \tfrac{d_U}{2} - \sum_l \text{AUC}(U^{(l)}) \;=\; -\sum_l \big(\text{AUC}(U^{(l)}) - \tfrac{1}{2}\big). \tag{45}$$

Combining. Substituting Eqs. (41) and (45),

$$\mathcal{U}_\alpha(\mathcal{M}^{\pi\text{-match}}) - \mathcal{U}_\alpha(\mathcal{M}^{\text{marg}}) \;=\; \Delta_N - \alpha\Delta_D \tag{46}$$

$$\geq\; -\sum_l \big(\text{AUC}(U^{(l)}) - \tfrac{1}{2}\big) + \alpha \sum_k \big(\text{AUC}(Z^{(k)}) - \tfrac{1}{2}\big), \tag{47}$$

which gives Eq. 17. □

B EXAMPLES SUPPORTING THE THEORY

The following example illustrates log-supermodularity in the case of a Gaussian distribution:

Example 2 (Log-supermodularity for Gaussians). *Suppose $U \mid Z = z \sim \mathcal{N}(\mu(z), \Sigma)$, where the mean $\mu(z)$ may depend on z but the covariance Σ does not. The log-density is*

$$\log P(u \mid z) = -\tfrac{1}{2}(u - \mu(z))^\top \Sigma^{-1}(u - \mu(z)) + \text{const}, \tag{48}$$

and a direct computation gives, for $k \neq l$,

$$\frac{\partial^2 \log P(u \mid z)}{\partial u^{(k)} \partial u^{(l)}} = -(\Sigma^{-1})_{kl}. \tag{49}$$

Log-supermodularity therefore reduces to the condition

$$(\Sigma^{-1})_{kl} \;\leq\; 0 \quad \textit{for all } k \neq l, \tag{50}$$

i.e., the off-diagonal entries of the precision matrix are non-positive, which is equivalent to all partial correlations being non-negative. □

The next example illustrates why log-supermodularity is necessary in Prop. 2:

Example 3 (Necessity of supermodularity). *Fix $Z = z$ and condition on it throughout. Consider $U = (U^{(1)}, U^{(2)}) \in \{0,1\}^2$, and the joint distribution*

$$P(U) = \begin{cases} 0.49 & \textit{if } u \in \{(1,0),(0,1)\} \\ 0.01 & \textit{if } u \in \{(0,0),(1,1)\}, \end{cases} \tag{51}$$

which is log-submodular, since $P(0,0)\,P(1,1) = 10^{-4} \ll 0.2401 = P(0,1)\,P(1,0)$. Let the propensity depend only on $U^{(1)}$,

$$\pi(u^{(1)}, u^{(2)}) = \begin{cases} 0.9 & \textit{if } u^{(1)} = 1 \\ 0.1 & \textit{if } u^{(1)} = 0, \end{cases} \tag{52}$$

which is non-decreasing in each coordinate of u. A direct computation gives

$$P(U \in A \mid X = 1) \approx 0.12, \qquad P(U \in A \mid X = 0) \approx 0.88, \tag{53}$$

for the upper set $A = \{u : u^{(2)} = 1\}$. Hence $P(U \mid X = 1) \not\succeq_{st} P(U \mid X = 0)$, and the dominance is in fact reversed along the $U^{(2)}$ coordinate. □

The intuition for the example can be described as follows. Conditioning on $X = 1$ pulls the distribution of $U^{(1)}$ upward, since π is increasing in $u^{(1)}$. The strong negative dependence between $U^{(1)}$ and $U^{(2)}$ then drags the distribution of $U^{(2)}$ downward, against the direction of the dominance claim. Log-supermodularity of $P(U \mid Z = z)$ rules out exactly this kind of trade-off, ensuring that the $\{X = 0 \to X = 1\}$-induced shift along one coordinate does not get reversed along another. In the context of our ICU example, this means that the assumption would fail if unobserved severity markers traded off against one another – e.g., if patients were sick along their respiratory axis *or* their cardiac axis, but rarely both. As argued above, this is unlikely to be true for our ICU setting, in which severity tends to cluster across organ systems.

For grounding Z-dominance of Prop. 4, we discuss the logistic-Gaussian setting:

Example 4 (Z-dominance in the logistic-Gaussian setting). *Let $U, Z \in \mathbb{R}$ be jointly Gaussian with unit variances and correlation ρ, and let*

$$P(X = 1 \mid Z, U) = \sigma(\alpha + \beta Z + \gamma U), \qquad \sigma(t) := (1 + e^{-t})^{-1}, \tag{54}$$

with $\beta, \gamma > 0$. Since U is univariate, the within-u inequality in Eq. 11 is vacuous and only the across-z-u inequality Eq. 11 must be checked. We work the $x = 0$ side. The cross derivative decomposes as

$$\frac{\partial^2 \log h^{(0)}(z,u)}{\partial u\, \partial z} = \underbrace{\frac{\partial^2 \log(1-\sigma)}{\partial u\, \partial z}}_{\textit{collider}} + \underbrace{\frac{\partial^2 \log P(u \mid z)}{\partial u\, \partial z}}_{\textit{bidirected}}. \tag{55}$$

A direct computation gives

$$\frac{\partial^2 \log(1-\sigma(\alpha+\beta z+\gamma u))}{\partial u\,\partial z} = -\beta\gamma \cdot \sigma(\alpha+\beta z+\gamma u)(1-\sigma(\alpha+\beta z+\gamma u)), \tag{56}$$

and, using $\log P(u \mid z) = -\frac{1}{2(1-\rho^2)}(u-\rho z)^2 + \text{const}$*,*

$$\frac{\partial^2 \log P(u \mid z)}{\partial u\,\partial z} = \frac{\rho}{1-\rho^2}. \tag{57}$$

The condition Eq. 11 therefore becomes

$$\frac{\rho}{1-\rho^2} \;\le\; \beta\gamma \cdot \sigma(\alpha+\beta z+\gamma u)(1-\sigma(\alpha+\beta z+\gamma u)). \tag{58}$$

We note that the same inequality is obtained on the $x = 1$ *side: in the logistic case* $\partial_z \partial_u \log\sigma = \partial_z \partial_u \log(1-\sigma) = -\beta\gamma\,\sigma(1-\sigma)$*, so both chains in Prop. 4 succeed under the same condition.*

Two regions of (u, z)*-space are illustrative. Near the decision boundary, where* $\alpha+\beta z+\gamma u \approx 0$ *and* $\sigma(1-\sigma) = 1/4$*, Eq. 58 reduces to* $\rho/(1-\rho^2) \le \beta\gamma/4$*, permitting* $\rho \lesssim 0.24$ *when* $\beta\gamma = 1$ *and* $\rho \lesssim 0.62$ *when* $\beta\gamma = 4$*. Further into the tails,* $\sigma(1-\sigma) \to 0$ *and the condition tightens to* $\rho \le 0$*.* □

The example shows that (iii) is genuinely weaker than $Z \perp\!\!\!\perp U$: arbitrary positive marginal correlation is permitted, provided the propensity coefficients are sufficiently large. It also makes the trade-off explicit: increasing $\beta\gamma$ relaxes Eq. 58, while in a multivariate extension increasing γ would make the within-U condition harder to satisfy: the regime that satisfies both is one of strong propensity responsiveness paired with strong intra-U marginal coupling.

The following examples illustrate how the chain Eq. 10 of Prop. 4 may fail:

Example 5 (Interaction in π at fixed X may break one chain even under $Z \perp\!\!\!\perp U$)**.** *Consider the SCM*

$$Z \sim \text{Bern}(0.5), \quad U \sim \text{Unif}[0,1], \quad Z \perp\!\!\!\perp U, \tag{59}$$

$$X \leftarrow \text{Bern}(\pi(Z,U)), \quad \pi(Z,U) = \begin{cases} U & \text{if } Z=0 \\ U & \text{if } Z=1, U<0.5 \\ 1 & \text{if } Z=1, U \ge 0.5. \end{cases} \tag{60}$$

A direct computation gives $\pi(z_0) = 0.5$*,* $\pi(z_1) = 0.625$*, and*

$$P(U > \tfrac{1}{2} \mid Z = z_1, X = 1) = 0.80, \qquad P(U > \tfrac{1}{2} \mid Z = z_0, X = 1) = 0.75, \tag{61}$$

$$P(U > \tfrac{1}{2} \mid Z = z_1, X = 0) = 0, \qquad P(U > \tfrac{1}{2} \mid Z = z_0, X = 0) = 0.25. \tag{62}$$

Interestingly, the two-step X *and* Z*-dominance still holds*

$$P(U > \tfrac{1}{2} \mid z_0, X = 1) = 0.75 \ge 0 = P(U > \tfrac{1}{2} \mid z_1, X = 0),$$

but only via the $X = 0$ *chain. The* $X = 1$ *chain fails (*$0.75 \not\ge 0.80$ *above), because the saturation of* π *in the* $(z_1, u \ge 0.5)$ *corner makes (iii) fail at* $X = 1$*. The example illustrates that (iii) is genuinely a per-chain assumption: an interaction in the propensity at fixed* X *can break one side without breaking the other, and Prop. 4 only needs one* $X = x$ *to work.* □

Our further example illustrates why strong positive correlation between Z and U may break the X and Z-dominance:

Example 6 (Strong positive Z-U correlation breaks both chains)**.** *Consider the SCM*

$$Z \sim \text{Bern}(0.5), \tag{63}$$

$$U \mid Z \sim \begin{cases} \text{Bern}(0.1) & \text{if } Z=0 \\ \text{Bern}(0.9) & \text{if } Z=1 \end{cases} \tag{64}$$

$$X \leftarrow \text{Bern}(\pi(Z,U)), \quad \pi(Z,U) = \begin{cases} 0.2 & Z=0, U=0 \\ 0.4 & Z=0, U=1 \\ 0.3 & Z=1, U=0 \\ 0.5 & Z=1, U=1, \end{cases} \tag{65}$$

where π is non-decreasing in each of Z, U. We compute $\pi(z_0) = 0.22$, $\pi(z_1) = 0.48$, and

$$P(U = 1 \mid Z = z_0, X = 1) \approx 0.18, \qquad P(U = 1 \mid Z = z_1, X = 0) \approx 0.87, \tag{66}$$

so Z-dominance fails: the smaller-Z treated unit is stochastically smaller along U than the larger-Z untreated unit. The strong positive Z-U marginal dependence concentrates joint mass on the $(0, 0)$ and $(1, 1)$ corners; (iii) fails for both $x = 0$ and $x = 1$, so neither chain is available. This is the second failure mode: if marginal Z-U dependence is too strong relative to the responsiveness of π, the bidirected force overwhelms the collider force, and dominance doesn't hold. □

C π-DOMINANCE

In this appendix, we discuss the π-dominance strategy. The same chain decomposition that produced Prop. 4 applies with the propensity score $\pi(Z)$ replacing the full vector Z, giving a π-dominance result that compares two propensity levels $p' < p$ at fixed treatment.

Proposition 5 (π-dominance)**.** *Under assumptions of Prop. 2, and additionally:*

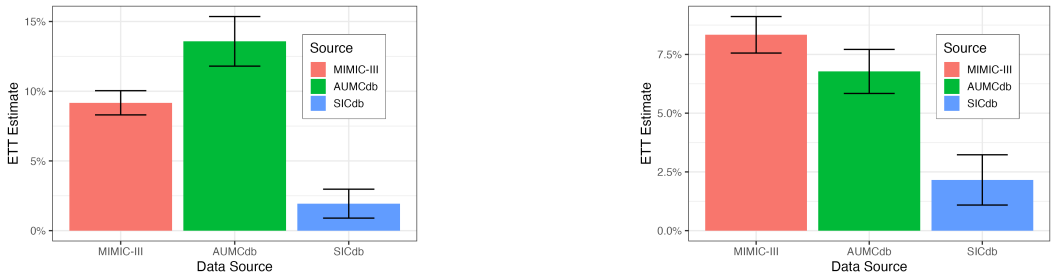
(iii') *For some $x \in \{0,1\}$, defining $h_\pi^{(x)}(p,u) \coloneqq P(U=u, \pi(Z)=p, X=x)$, $\log h_\pi^{(x)}$ is twice differentiable and strictly positive, with*

$$\frac{\partial^2 \log h_\pi^{(x)}}{\partial u^{(j)}\,\partial u^{(k)}}(p,u) \;\geq\; 0 \quad \textit{for all } j \neq k, \qquad \frac{\partial^2 \log h_\pi^{(x)}}{\partial p\,\partial u^{(j)}}(p,u) \;\leq\; 0 \quad \textit{for all } j. \tag{67}$$

Then, for $p' < p$,

$$P(U \mid \pi(Z)=p', X=1) \;\succeq_{st}\; P(U \mid \pi(Z)=p, X=0). \tag{68}$$

The proof is identical to that of Prop. 4, with the scalar p replacing z throughout: Prop. 3 provides the π-matching link of the chain, and (iii') provides the π-shift link via Karlin–Rinott. The collider/bidirected interpretation of Prop. 4 carries over: the within-U inequality again asks that intra-U marginal clustering survives the collider re-weighting, while the across-π-U inequality asks that the collider channel overturn the marginal dependence between $\pi(Z)$ and U. The conceptual gain over Z-dominance is that the across-direction inequality is now a single one-dimensional cross-partial, regardless of $\dim(Z)$: the propensity score concentrates the structural assumption along the single dimension that actually drives treatment assignment. The within-U inequality, by contrast, is not strictly weaker than its Z-dominance counterpart, since $h_\pi^{(x)}$ is a marginalization of $h^{(x)}$ over the level set $\{z : \pi(z) = p\}$, and log-supermodularity is not preserved under marginalization in dimensions higher than two; the conditions (iii) and (iii') of Props. 4 and 5 are therefore similar in spirit but logically independent (one does not imply the other).



(a) Mechanical ventilation ETTs.

(b) Vasopressor initiation ETTs.

Figure 5: Estimated $\widehat{\text{ETT}}^{bd}$ on in-hospital mortality across three ICU databases (MIMIC-III, AUMCdb, SICdb), for two distinct ICU interventions. Error bars indicate 95% confidence intervals.

D Extended Discussion of Ex. 1

In Ex. 1 of Sec. 1, we argued that the positive $\widehat{\text{ETT}}^{bd}$ of mechanical ventilation on in-hospital mortality is a near-certain signature of unobserved confounding. In this appendix, we provide an extended discussion of this claim, making each step of the reasoning explicit.

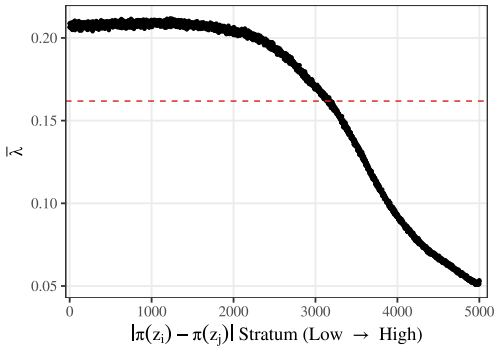
(i) The estimand is the ETT, and not the ATE. The choice of the ETT as an estimand in Eq. 1, rather than the average treatment effect (ATE), is deliberate. Ventilation is a treatment whose individual-level effect may be genuinely heterogeneous: it supports oxygenation, but carries risk through sedation, duration of exposure, and adverse ventilator-associated events. Therefore, the average treatment effect $\mathbb{E}[Y_{x_1} - Y_{x_0}]$, which averages the treatment effect *over the entire population*, may in fact show harm. The reason is that the ATE also includes patients for whom physicians never consider for ventilation. Therefore, a positive ATE (indicating overall harm) cannot be ruled out apriori, since it may be the case that treating every patient is worse than treating none of the patients.

The reasoning for the ETT quantity is different, however. The ETT averages the causal effect over the subpopulation $\{X = x_1\}$, that is, over *exactly those patients the attending physician decided to treat*. This means that the physician, at the bedside and with access to the full clinical picture, assessed that the treatment was likely more beneficial than harmful. A genuinely harmful ETT estimate would therefore imply that removing all ventilation devices from the ICU would improve outcomes *relative to current practice*. This claim is extremely unlikely to be true.

(ii) The magnitude of the estimates is large relative to baseline risk. The second point worth considering relates to the effect sizes. Baseline in-hospital mortality across the three databases lies between 7.4% and 8.4%, so that the estimated harmful effects in Fig. 5a are very large in relative terms. Mechanisms such as complications, prolonged ventilation, sedation, or ventilator-associated events act on a subset of the treated and over a limited window. For them to account for harmful effects of this size (in some datasets nearly doubling the risk of mortality), their aggregate harm would need to exceed a substantial fraction of the total baseline mortality of the ICU population, while conferring essentially no benefit on the remaining treated patients. Excess risk of this magnitude is not what the clinical literature on the treatment would lead one to expect, and its size is itself informative: it points towards a UC bias (treated and untreated patients are not comparable, due to unobserved factors) rather than a harmful causal effect.

(iii) The pattern is not specific to ventilation. Finally, one may hypothesize that factors related to timing or the specific version of the ventilation treatment may result in erroneous or biased effect estimates. Crucially, we remark that ventilation is just one of the many ICU treatment showing the same behavior. Almost every treatment displays the same signature. A particularly instructive case is the initiation of vasopressors, shown in Fig. 5b, with $\widehat{\text{ETT}}^{bd}$ equal to 7.2% ($\pm 0.8\%$) in AUMCdb, 8.5% ($\pm 0.9\%$) in MIMIC-III, and 2.1% ($\pm 0.8\%$) in SICdb. These findings are very similar to those obtained for ventilation. Vasopressors, however, are typically given to patients whose hemodynamic status is nearly incompatible with life, making a genuinely harmful ETT even harder to explain than

in the ventilation case. Therefore, it is again more plausible that the physician observes signs of deterioration that are never recorded in the EHR, and allocates treatment on their basis, resulting in biased ETT estimated due to existence of UCs.



Pair budget B	Spearman mean [95% CI]	Prec@10 mean [95% CI]
50	0.55 [0.52, 0.58]	0.68 [0.66, 0.70]
500	0.68 [0.67, 0.69]	0.74 [0.73, 0.75]
2000	0.69 [0.69, 0.70]	0.75 [0.74, 0.76]

Figure 6: Additional semi-synthetic results on MIMIC-III-SeS. *Left*: mean $\lambda(i, j)$ across strata of $|\hat{\pi}(Z_i) - \hat{\pi}(Z_j)|$, from low to high; the dashed horizontal line is the overall mean success rate, corresponding to the success rate of $\mathcal{M}^{\text{marg}}$. *Right*: recovery of the ground-truth candidate ordering under $\mathcal{M}^{\pi\text{-match}}$ as the pair budget B increases.

E Additional semi-synthetic experiments

In this appendix, we provide additional details and results for the semi-synthetic experiments in Sec. 3.2. First, we examine the relationship between propensity-score similarity $|\hat{\pi}(Z_i) - \hat{\pi}(Z_j)|$ and the pair (i, j) success probability $\lambda(i, j)$ (probability of extracting a genuine unobserved confounder). This analysis demonstrates how well the $\mathcal{M}^{\pi\text{-match}}$ strategy is expected to work. Second, we investigate whether the matching strategies can distinguish a priority among UCs, that is, whether they give higher importance to stronger UCs.

Pair-level success and propensity-score similarity. For each pair (i, j), we compute the absolute difference in observed treatment propensity $|\hat{\pi}(Z_i) - \hat{\pi}(Z_j)|$. We then stratify pairs into bins, from low to high absolute propensity difference. We compute the mean pair success probability $\lambda(i, j)$ within each bin. Fig. 6 (left) shows the values of the average $\bar{\lambda}$ in each bin. The average $\lambda(i, j)$ is highest among pairs with similar estimated propensities, and decreases as the propensity gap widens. The dashed horizontal line gives the overall mean of $\lambda(i, j)$ across all pairs (i, j), corresponding to the success rate obtained under marginal matching $\mathcal{M}^{\text{marg}}$. Therefore, pairs selected to have similar $\hat{\pi}(Z)$ are systematically more informative than marginally sampled pairs, empirically illustrating the mechanism behind Thm. 3. Importantly, here, the joint (Z, U) distribution is inherited directly from real-world clinical data.

Prioritizing UCs. The original MIMIC-III-SeS construction contains $d_U = 10$ binary unobserved confounders U, and additional $d_{U^{\text{fp}}} = 10$ unobserved false positives U^{fp}. Each U affects both treatment and outcome, while U^{fp} influences neither the treatment nor the outcome. For a strategy M and pair budget B, let $\{(i_b, j_b)\}_{b=1}^{B}$ denote its selected treated-untreated pairs. We say that a candidate variable $U^{(l)}$ receives a hit whenever

$$U_{i_b}^{(l)} > U_{j_b}^{(l)}, \tag{69}$$

that is, whenever the variable's value is larger for the treated unit than the untreated one. Its cumulative hit score is given by

$$H_l^M(B) \coloneqq \sum_{b=1}^{B} \mathbb{1}\left\{U_{i_b}^{(l)} > U_{j_b}^{(l)}\right\}. \tag{70}$$

Candidates are ranked in decreasing order of $H_l^M(B)$. Notice that false positives candidates in U^{fp} can also receive hits, but a successful design should make the genuine treatment drivers U appear systematically more often compared to the false positives U^{fp}.

To assess whether this is the case, we define the ground-truth importance of candidate $U^{(l)}$ by

$$I_l \coloneqq \hat{p}_l(1 - \hat{p}_l)|\gamma_X^{(l)}|, \qquad \hat{p}_l \coloneqq \frac{1}{n}\sum_{i=1}^{n} U_i^{(l)}.$$

The importance score is proportional to the strength of the confounders influence on X, $|\gamma_X^{(l)}|$, and also the variable's amount of variation in the population $\hat{p}_l(1-\hat{p}_l)$, where $\hat{p}_l$ is the prevalence of $U^{(l)}$. The false-positive candidates have $I_l = 0$, since $\gamma_X^{(l)} = 0$. At each budget B, we compare the ranking induced by $H_l^M(B)$ with this ground-truth importance score, using Spearman rank correlation. We additionally report the fraction of the ten highest-ranked candidates that are genuine confounders, defined as

$$\mathrm{Prec@10} = \frac{|U \cap U^M_{\text{Top-10}}(B)|}{10}. \tag{71}$$

We report mean values and CIs over 100 random repetitions.

As shown in Fig. 6 (right), prioritization improves as the comparison budget B increases. Under $\mathcal{M}^{\pi\text{-match}}$, the mean Spearman correlation with the ground-truth ordering increases from 0.55 at $B = 50$ to 0.69 at $B = 2000$, while $\mathrm{Prec@10}$ increases from 0.68 to 0.75. Thus, repeated matched comparisons do more than identify individual candidate variables: their frequency provides information about which candidates are most relevant to the treatment decision.

CUI C0#	Concept	G	CUI C0#	Concept	G
32227	Pleural effusion	A	10054	Coronary athero.	D
19080	Hemorrhage	C	04144	Atelectasis	A
32285	Pneumonia	A	16658	Fracture	C
13604	Edema	E	09450	Commun. disease	B
13404	Dyspnea	A	332448	Infiltration	A
30193	Pain	E	27497	Nausea	E
151699	Intracranial hem.	C	242184	Hypoxia	A
20538	Hypertensive dis.	D	31039	Pericardial eff.	D
15967	Fever	B	39239	Sinus tachycardia	B
34063	Pulm. edema	A	32326	Pneumothorax	A

Table 1: Top 20 discovered confounders (ranked by frequency) on real data via ε_{abl} with $B = 2{,}000$ pairs via $\mathcal{M}^{\pi\text{-match}}$. Groups: (**A**) pulmonary impairment/injury; (**B**) infection/SIRS; (**C**) hemorrhage/trauma; (**D**) cardiac; (**E**) non-specific.

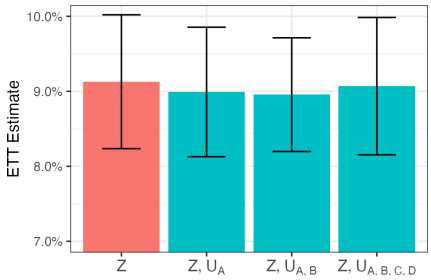


Figure 7: ETT estimates on MIMIC-III with adjustment sets Z, $\{Z, U_A\}$, $\{Z, U_{A,B}\}$, and $\{Z, U_{A,B,C,D}\}$, using causal forests with 100 bootstrap replicates (95% CIs reported).

F Mechanical Ventilation in MIMIC-III Real Data

We now apply our framework to the ICU setting that motivated the paper (Ex. 1). Recall that the EHR-based ETT estimate of mechanical ventilation on in-hospital mortality is strongly above zero in every database (Fig. 1), a near-certain signature of unobserved confounding. We perform confounder detection to elicit confounder candidates for MIMIC-III, then re-estimate the ETT.

We first fit a BERT-based estimator $\hat{\pi}_{\text{BERT}}$ that predicts $P(X = 1 \mid Z, T)$ (mechanical ventilation) from the physiological covariates Z together with the clinical notes T, using the training subset of the data. We further fit a predictor $\hat{\pi}_{\text{xgb}}(Z)$ for predicting X just based on the covariates Z using `xgboost` (Chen & Guestrin, 2016), and such predicted probabilities are used for the $\mathcal{M}^{\pi\text{-match}}$ strategy. On a held-out set, we select $B = 2{,}000$ treated-control pairs (i, j) using $\mathcal{M}^{\pi\text{-match}}$, ensuring that any unit i or j can appear up to three times. For each selected pair, we run ablation-based extraction ε_{abl}. We extract the UMLS concepts present in patient i's notes but not in j's (set denoted by $\mathcal{C}_i \setminus \mathcal{C}_j$), and for each concept $c \in \mathcal{C}_i \setminus \mathcal{C}_j$ we remove its mentions from patient i's notes and recompute the propensity for treatment labeled $\hat{\pi}(z_i, t_i^{-c})$. We define the concept impact as $\Delta_c = \hat{\pi}_{\text{BERT}}(z_i, t_i) - \hat{\pi}_{\text{BERT}}(z_i, t_i^{-c})$. Concepts with $\Delta_c > 1\%$ are recorded as candidate confounders.

Aggregating across the 2,000 pairs, we report the 20 most frequently discovered concepts in Tab. 1. We classify the concepts into groups, based on their clinical interpretation. Group **(A)** consists of concepts with direct indication for mechanical ventilation, related to pulmonary injury or impairment. Group **(B)** consists of concepts (fever, tachycardia, communicable disease) related to Systemic Inflammatory Response Syndrome (SIRS (Bone et al., 1992)), which is a known pathway that drives ventilation via septic respiratory failure (Esteban et al., 2000). Group **(C)** is related to hemorrhage and trauma, which drive MV via shock and airway protection. Group **(D)** contains cardiac comorbidities, which act indirectly via hemodynamic instability or background risk (Esteban et al., 2000). Finally, Group **(E)** contains non-specific correlates of illness severity, which are less likely to be recognized as drivers of ventilation decisions. In summary, Group (A) has very high clinical plausibility in terms of confounding MV, while Groups (B)-(D) have high plausibility. Therefore, 17/20 detected confounders are clinically highly plausible.

Effect on ETT estimation. We next assess whether adjusting for the detected confounders impacts the ETT estimate comparing to using Z only. We compute the ETT on the entire MIMIC-III data using causal forests (Athey & Wager, 2019) with four adjustment sets: Z, (Z, U_A), $(Z, U_{A,B})$, and $(Z, U_{A,B,C,D})$, with U_G encoding binary indicators for the concepts in groups G. Fig. 7 shows the effect estimates for different adjustment sets. The figure indicates that incorporating U_A, U_B reduces the ETT slightly, while incorporating $U_{A,B,C,D}$ pulls the estimate back to a larger value. However, none of the effect estimates differ to a statistically significant level. While most of the elicited confounders are clinically plausible, accounting for their impact does not reduce the ETT towards 0. We hypothesize that this is due to the retrospective and partial nature of the text data: even highly relevant concepts such as pleural effusion are documented for only a minority of patients, so the binary indicators we construct may be noisy proxies rather than the underlying confounder itself.