Annotation-Assisted Learning of Treatment Policies From Multimodal Electronic Health Records

Henri Arno HENRI.ARNO@UGENT.BE
Ghent University - imec, Ghent, Belgium

Thomas Demeester THOMAS.DEMEESTER@UGENT.BE
Ghent University - imec, Ghent, Belgium

Abstract

We study how to learn treatment policies from multimodal electronic health records (EHRs) that consist of tabular data and clinical text. These policies can help physicians make better treatment decisions and allocate healthcare resources more efficiently. Causal policy learning methods prioritize patients with the largest expected treatment benefit. Yet, existing estimators are designed for tabular covariates under causal assumptions that may be hard to justify in the multimodal setting. A pragmatic alternative is to apply causal estimators directly to multimodal representations, but this can produce biased treatment effect estimates when the representations do not preserve the relevant confounding information. As a result, predictive models of baseline risk are commonly used in practice to guide treatment decisions, although they are not designed to identify which patients benefit most from treatment. We propose AACE (Annotation-Assisted Coarsened Effects), an *annotation-assisted* approach to causal policy learning for multimodal EHRs. The method uses expert-provided annotations during training to support confounding adjustment, and then predicts treatment benefit from only multimodal representations at inference. We show that the proposed method achieves strong empirical performance across synthetic, semi-synthetic, and real-world EHR datasets, outperforming risk-based and representation-based causal baselines, and offering practical insights for applying causal machine learning in clinical practice.

1. Introduction

Machine learning (ML) is increasingly being used in healthcare to support clinical decision-making (Esteva et al., 2019; Rajkomar et al., 2019). A central problem in this context is learning from observational data which patients should receive treatment, supporting more targeted and efficient allocation of clinical resources (Athey and Wager, 2021; Nie et al., 2021). Electronic health records (EHRs) are widely available in modern healthcare systems and capture rich, patient-specific information, making them well suited for learning data-driven treatment policies (Tang et al., 2024; Shickel et al., 2017). However, these records are typically **multimodal**, combining tabular data and clinical text. In this work, we study how to learn *effective* and *reliable* treatment policies from such multimodal EHR data.

One strategy to tackle this problem is to directly estimate treatment effects and prioritize patients with the largest expected benefit (Feuerriegel et al., 2024; Bica et al., 2021). By modeling how outcomes change under different treatment assignments, this strategy aligns the learning objective with the goal of making well-targeted treatment decisions. However, existing estimators are designed for *tabular* covariates under causal identification assumptions, such as unconfoundedness. These assumptions can be difficult to justify in

multimodal EHRs, where patient information is distributed across tabular data and clinical text. A pragmatic alternative is therefore to apply causal estimators directly to representations of the multimodal data, implicitly assuming that they are sufficient for confounding adjustment. However, when relevant confounding information is missing from these representations, this leads to **biased** treatment effect estimates.

Because of these challenges, a more common strategy in practice is *predictive* modeling of the clinical outcome in absence of treatment (Van Smeden et al., 2021; Goldstein et al., 2016). The resulting predictions form an estimate of the patients' so-called *baseline risk* and treatment can be prioritized for those most at risk. This approach is appealing because predictive models can easily handle multimodal data, and richer data often translates into higher predictive accuracy. However, as these models do not target treatment effects directly, they may fail to identify the patients who would benefit most, especially when baseline risk and treatment benefit are not well aligned. For instance, a patient may be at high baseline risk yet benefit little from treatment, while another with lower risk may benefit more, leading to **suboptimal** treatment allocation.

Proposed Method. We extend the causal strategy to the multimodal setting with an *annotation-assisted* approach, which we refer to as AACE (Annotation-Assisted Coarsened Effects). The method builds on the doubly robust learner (Kennedy, 2023) to estimate *coarsened* treatment effects, which quantify the expected treatment benefit given the information contained in the multimodal representations. When relevant confounders are expressed in clinical text, but not entirely preserved in these representations, reliable treatment effect estimation requires additional training-time covariates that are sufficient for confounding adjustment. In AACE, expert-provided annotations serve as a pragmatic way to obtain such covariates during training. At inference time, the model relies *only* on the multimodal representations to assign treatment.

Main Results. Our experiments show that the proposed method can learn effective treatment policies from multimodal EHR data given suitable supervision. On synthetic and semi-synthetic benchmarks, AACE *outperforms both risk-based and representation-based causal baselines*. Additional robustness analyses show that the method remains effective under moderately imperfect annotations and representations. On real-world clinical data, the different methods produce substantially different treatment assignments, showing that the choice of method directly affects which patients are prioritized for treatment.

Contributions. Together, these findings motivate the following three contributions:

- We study the problem of learning treatment policies from multimodal EHR data, where patient information is distributed across tabular data and clinical text, and standard causal assumptions may not hold for the data representations directly.
- We introduce AACE, an *annotation-assisted* method for causal policy learning on multimodal data, which uses expert annotations during training to support confounding adjustment while relying only on multimodal representations at inference.

- We empirically evaluate the proposed method across synthetic, semi-synthetic, and real-world EHR datasets. The results show that AACE (i) outperforms both risk-based and representation-based causal baselines, (ii) remains effective under moderately imperfect annotations and representations, and (iii) generates materially different treatment assignments on real-world data.

Generalizable Insights about Machine Learning in the Context of Healthcare

Our findings suggest several broader lessons for machine learning in healthcare. In multimodal settings, predictive models of baseline risk are not necessarily appropriate for treatment assignment, because risk and treatment benefit do not necessarily align. At the same time, directly applying causal estimators to learned representations can be problematic when those representations do not capture all relevant confounders, leading to biased effect estimates. Expert-provided annotations offer a *practical* way to support confounding adjustment during training, helping causal methods better match the realities of multimodal clinical data. More broadly, the choice of modeling strategy can directly affect which patients are prioritized for treatment, making the method choice matter in practice.

2. Background and Related Work

We summarize the foundations of learning treatment policies from observational data and review related work on the main strategies for this task.

2.1. Learning Treatment Policies From Observational Data

Consider a setting where each individual i in the population is characterized by covariates $X_i \in \mathcal{X}$, receives a binary treatment $T_i \in \{0, 1\}$, and has an associated continuous outcome $Y_i \in \mathbb{R}$. In the Rubin-Neyman potential outcomes framework (Rubin, 2005), each individual also has two potential outcomes, $Y_i(1)$ and $Y_i(0)$, where $Y_i(t)$ denotes the outcome that we would observe if individual i were assigned treatment $T_i = t$. In practice, only one of the two potential outcomes is observed, corresponding to the treatment that was actually received.

A *treatment policy* π is a rule that maps individual covariates to a treatment decision $\pi(x) : \mathcal{X} \to \{0, 1\}$. In the potential outcomes framework, the performance of a policy can be quantified by its *policy value*,

$$V(\pi) = \mathbb{E}\big[\, Y\big(\pi(X)\big)\big] = \mathbb{E}\big[\, \pi(X)\, Y(1) + \big(1 - \pi(X)\big)\, Y(0)\big], \tag{1}$$

which represents the mean outcome in the population if treatment were assigned according to π. The optimal policy π^* is the one that maximizes the policy value (assuming larger outcomes are better). In many applications, however, treatment resources are limited and only k individuals in a population *can* be treated. In such constrained settings, the optimal policy maximizes the policy value while satisfying the resource constraint.

In practice, we aim to learn treatment policies from observed data $(X, T, Y) \sim P$, where the joint distribution factorizes as $P(X, T, Y) = P(X)\, P(T \mid X)\, P(Y \mid X, T)$. Here, $P(T \mid X)$ represents the existing, possibly suboptimal, *observational* policy $\pi_{\text{obs}}(x)$. The goal of treatment policy learning is to use this data to estimate a new policy $\hat{\pi}$ that achieves a higher policy value than π_{obs} and approximates π^*.

2.2. Causal Treatment Policies

Causal strategies learn treatment policies by modeling how outcomes would change under different treatment assignments. Within this framework, two main approaches have emerged. First, in *direct policy learning*, the treatment policy is learned directly by optimizing an empirical estimate of the policy value (Qian and Murphy, 2011; Kallus, 2018; Athey and Wager, 2021). Second, in *CATE-based policy learning*, the conditional average treatment effect (CATE) is estimated first, and individuals whose estimated benefit exceeds a threshold are treated (Arno et al., 2026; Frauen et al., 2025; Fernández-Loría and Provost, 2022). We focus on this latter family because it forms the basis of our proposed method.

The conditional average treatment effect (CATE) quantifies the expected benefit of treatment for an individual with covariates $X = x$,

$$\tau^x(x) = \mathbb{E}\big[Y(1) - Y(0) \mid X = x\big]. \tag{2}$$

Because the potential outcomes are not directly observable, we rely on the following standard causal inference assumptions for the identification of the CATE:

- **Assumption 1.1 (consistency).** The observed outcome equals the potential outcome under the received treatment: $Y(t) = Y$ whenever $T = t$.
- **Assumption 1.2 (positivity).** Each individual has a non-zero probability of receiving either treatment: $0 < e(x) < 1$, where $e(x) = \pi_{\text{obs}}(x) = P(T = 1 \mid X = x)$ is the *propensity score*.
- **Assumption 1.3 (unconfoundedness).** There are no unmeasured confounders: $Y(t) \perp\!\!\!\perp T \mid X$.

Under these assumptions, the CATE is identified, meaning that it can be estimated from observational data, and may be expressed as

$$\tau^x(x) = \mathbb{E}\big[Y \mid T{=}1, X{=}x\big] - \mathbb{E}\big[Y \mid T{=}0, X{=}x\big] = \mu_1(x) - \mu_0(x), \tag{3}$$

where $\mu_t(x) = \mathbb{E}\big[Y \mid T = t, X = x\big]$ denotes the conditional mean outcome under treatment $t \in \{0, 1\}$, commonly referred to as the *response surfaces*. Together with the propensity score, these quantities form the *nuisance functions*, collectively referred to as $\eta = (e, \mu_1, \mu_0)$.

There is a rich literature on estimating the CATE from observational data. Existing machine learning models have been adapted for this purpose, including Gaussian processes, random forests, and generative adversarial networks (Alaa and van der Schaar, 2017; Wager and Athey, 2018; Yoon et al., 2018). More recently, meta-learners have gained popularity as general-purpose procedures for CATE estimation that can be instantiated with standard supervised learning models (Künzel et al., 2019; Nie and Wager, 2020; Kennedy, 2023). In their standard form, however, these estimators are typically developed for *tabular* covariates that are directly observed, and satisfy the above assumptions (1.1 – 1.3). In *multimodal* clinical data, these assumptions are difficult to satisfy in practice, because relevant confounding information may be distributed across tabular variables and clinical text, or may be only imperfectly preserved in the data representations.

2.3. Representation-Based CATE Estimation

Several authors have proposed neural network architectures that learn intermediate representations of *tabular* covariates for CATE estimation under the standard causal inference assumptions. Neural architectures such as TARNet learn a shared representation of the covariates and use treatment-specific heads to estimate the potential outcomes (Johansson et al., 2016; Shalit et al., 2017). DragonNet extends this design with an additional head for the propensity score, while subsequent work has explored broader architectural variants and inductive biases for representation-based treatment effect estimation (Shi et al., 2019; Curth and van der Schaar, 2021). However, when these learned representations fail to preserve relevant confounding information, the resulting treatment effect estimates can be biased (Melnychuk et al., 2024).

Recent work has also studied treatment effect estimation from high-dimensional and multimodal data, particularly text (Ma et al., 2025; Egami et al., 2022; Daoud et al., 2022; Gultchin et al., 2021; Veitch et al., 2020; Keith et al., 2020; Wood-Doughty et al., 2018), and more recently images, and combinations of both (Zhu et al., 2025; Klaassen et al., 2024; Jerzak et al., 2023; Sanchez and Tsaftaris, 2022). In these settings, the learned representations are typically assumed to be *sufficient for confounding adjustment*. In contrast, we argue that this assumption can be hard to satisfy in multimodal EHRs, and instead propose *annotation-assisted* learning as a more practical alternative. This is closely related to learning with privileged information, where richer variables are available during training than at inference (Lopez-Paz et al., 2016; Makar et al., 2019). AACE applies this idea to multimodal confounding, using expert annotations for confounding adjustment before learning treatment effects over inference-time representations.

2.4. Risk-Based Treatment Policies

An alternative strategy for learning treatment policies is to rely on *predictive* or *risk-based* models that estimate an individual's expected outcome in the absence of treatment, $\mu_0(x) = \mathbb{E}\big[Y \mid T = 0, X = x\big]$. This quantity can be interpreted as an individual's baseline risk, and treatment may then be prioritized for those with the highest predicted risk. Recent work has compared such strategies to causal methods (Fernández-Loría and Provost, 2022; Fernández-Loría et al., 2023; Athey et al., 2025), showing that risk-based policies *can* perform competitively, for instance when treatment effects and baseline risk are strongly correlated, or when treatment effects are difficult to estimate in small samples.

This strategy is particularly appealing in the multimodal setting, because it does not rely on causal identification assumptions, and predictive accuracy often improves with richer data. For instance, clinical outcome prediction can improve substantially when incorporating the multiple modalities in electronic health records (Liu et al., 2018; Yang and Wu, 2021). However, while such models may predict outcomes well, our goal is to learn *effective treatment policies* rather than to optimize predictive accuracy alone.

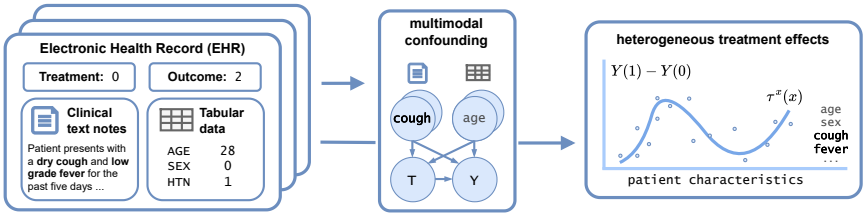

Figure 1: Overview of the problem setting. **Left:** Electronic health records combine tabular data and clinical text: key confounders (e.g., *cough* and *fever*) may appear only in text. **Middle:** Multimodal confounding influences both treatment T and outcome Y. **Right:** Heterogeneity in the conditional average treatment effect $\tau^x(x)$ across patient characteristics.

3. Problem Setting

In this section, we describe the structure of the multimodal electronic health records considered in this work and how expert annotations can address the confounding challenges they introduce.

3.1. Multimodal Confounding

Electronic health records (EHRs) are routinely collected in healthcare systems and capture detailed patient information, making them a natural choice for learning treatment policies from real-world data. These records typically combine tabular variables, such as demographics and lab measurements, with clinical text, such as nursing notes describing patient symptoms or radiology reports summarizing imaging findings (Johnson et al., 2016). Because of this multimodality, clinical factors that influence both treatment decisions and patient outcomes may appear only in the text, as illustrated in Figure 1. Estimating treatment effects from such data therefore requires adjustment for confounders that may be distributed across modalities.

3.2. Annotation-Assisted Confounding Adjustment

The clinical notes in an EHR can be encoded using a pre-trained text model, such as ModernBERT (Warner et al., 2024), to obtain an embedding. These embeddings, once concatenated with the tabular variables, form a multimodal representation ϕ of a record. However, if the text captures confounding factors only partially, or if the encoder does not preserve this information entirely, *unconfoundedness does not hold with respect to* ϕ. In that case, treatment effects are **not** identified from observational data (ϕ, T, Y) alone.

Reliable treatment effect estimation in this setting therefore requires training-time covariates that are sufficient for confounding adjustment. In multimodal EHRs, when clinically relevant confounders are expressed in the text, but not reliably preserved in ϕ, some explicit representation of these variables is **necessary** during training. In this work, we argue that *expert-provided annotations* offer a pragmatic way to obtain such covariates. Concretely, the annotations should capture clinically relevant factors, documented in the notes, that influence both treatment decisions and patient outcomes. In the example of Figure 1, this would correspond to annotating symptom variables such as *cough* or *fever*.

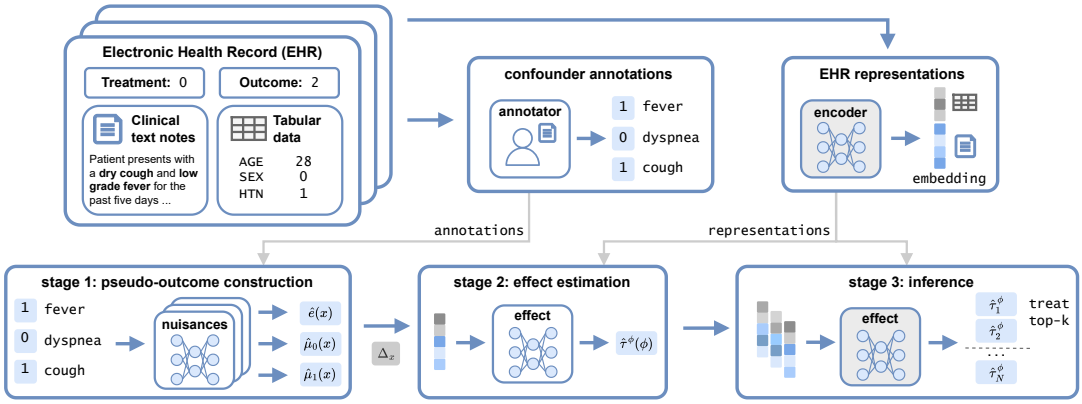

Figure 2: Overview of AACE. **Top:** Each multimodal electronic health record is (1) annotated during training to capture confounders and (2) encoded into a multimodal representation using a pre-trained model. **Bottom:** AACE proceeds in three stages: (1) nuisance estimation and pseudo-outcome construction using the causally sufficient training covariates, (2) treatment effect estimation from multimodal representations, and (3) inference-time treatment assignment.

Which covariates need to be annotated is application-dependent and should be guided by domain expertise. We therefore deliberately do not prescribe a fixed annotation protocol. In practice, such annotations may be obtained through a one-time annotation effort on top of routine clinical documentation, or through retrospective expert review. Appendix A provides a concrete example of how such annotations may be obtained in practice.

While this introduces an additional annotation cost, we view it as a practical compromise: rather than assuming that the representation alone preserves all confounding information, we explicitly collect the variables needed to support reliable confounding adjustment. Formally, we let X denote the training-time covariates that are assumed sufficient to adjust for confounding, i.e., $Y(t) \perp\!\!\!\perp T \mid X$, together with consistency and positivity. In our setting, X comprises the tabular variables in the EHR, together with the expert-provided annotations of confounders expressed in the text. The training data therefore consists of (ϕ, X, T, Y), whereas at inference, only the representation ϕ is available.

4. Proposed Method: AACE

We now describe AACE (Annotation-Assisted Coarsened Effects), our *annotation-assisted* method for treatment policy learning from multimodal EHR data. The method uses expert-provided annotations to support confounding adjustment during training, and then learns to predict treatment benefit from multimodal representations alone at inference time.

4.1. Coarsened Treatment Effects.

Because treatment decisions at inference can only depend on the multimodal representations, we define the target estimand as the expected treatment effect conditional on ϕ,

$$\tau^{\phi}(\phi) := \mathbb{E}\big[Y(1) - Y(0) \mid \phi\big] = \int \tau^{x}(x)\, p(x \mid \phi)\, dx. \tag{4}$$

Thus, $\tau^{\phi}(\phi)$ averages the treatment effects $\tau^{x}(x)$ over the patient subgroups whose covariates are compatible with the representation. It interpolates between the average treatment effect (ATE) when ϕ is uninformative about X, and the conditional average treatment effect (CATE) when ϕ fully determines X. The equality in (4) follows from the law of iterated expectations, under the assumption that the representations provide no additional information about the potential outcomes beyond the covariates, i.e., $Y(t) \perp\!\!\!\perp \phi \mid X$ for $t \in \{0, 1\}$. In our setting, this is natural because ϕ is derived from clinical data that encode the underlying covariates.

Treatment Policies Based on Coarsened Effects. As discussed in Section 2.2, CATE-based policies allocate treatment to the individuals with the highest expected treatment benefit. When the multimodal representations lose confounding information, the coarsened effects $\tau^{\phi}(\phi)$ may differ from the true effects $\tau^{x}(x)$. Nonetheless, the implied treatment policy can remain *optimal* when the bias induced by coarsening is small relative to the treatment effect margins. To see this, let $\delta_i = |\tau^{x}(x_i) - \tau^{\phi}(\phi_i)|$ denote the coarsening bias for individual i. For a fixed treatment budget k, let γ_k denote the treatment effect margin at the decision threshold. If individuals are ordered by decreasing (true) treatment effects, then γ_k is the difference between the k-th and $(k+1)$-th largest treatment effects.

Proposition 1 *For a fixed treatment budget k, if $\max_i \delta_i < \frac{\gamma_k}{2}$, then ranking individuals by $\tau^{\phi}(\phi)$ yields the same top-k treatment assignment as ranking them by $\tau^{x}(x)$.*

Thus, representation-based coarsening can distort effect values without affecting treatment assignment, as long as no individuals are swapped across the treatment decision boundary. A proof is given in Appendix B. In practice, the condition of Proposition 1 cannot be verified directly, since the true treatment effects are unknown, much like the standard causal identification assumptions cannot be verified from observational data alone (Assumptions 1.1 – 1.3). However, if the annotated covariates can be recovered reasonably well from the representations in the training data, this is a strong indication that the coarsening bias is limited and unlikely to alter treatment decisions.

4.2. Estimation with Annotation-Assisted Learning.

To estimate the coarsened treatment effects, AACE proceeds in three stages (cf. Figure 2).

Stage 1: Pseudo-Outcome Construction. We begin by constructing pseudo-outcomes following the doubly robust (DR) learner, a standard meta-learner for treatment effect estimation (Kennedy, 2023). This step uses the additional training-time covariates X to estimate the nuisance functions: the propensity score $\hat{e}(x)$ and the response models $\hat{\mu}_t(x)$

for $t \in \{0, 1\}$. The resulting pseudo-outcome is

$$\Delta_x = \hat{\mu}_1(x) - \hat{\mu}_0(x) + \frac{T - \hat{e}(x)}{\hat{e}(x)\big(1 - \hat{e}(x)\big)}\big(Y - \hat{\mu}_T(x)\big). \tag{5}$$

By the doubly robust property, if either the propensity model or the response models are correctly specified, then $\mathbb{E}[\Delta_x \mid X = x] = \tau^x(x)$. Thus, Δ_x provides an unbiased signal for the treatment effect, given the causally sufficient training-time covariates X.

Stage 2: Effect Estimation From Multimodal Representations. In the second stage, we regress the pseudo-outcomes Δ_x on the multimodal representations ϕ. By the law of iterated expectations,

$$\mathbb{E}\big[\Delta_x \mid \phi\big] = \mathbb{E}\big[\mathbb{E}\big[\Delta_x \mid X\big] \mid \phi\big] = \mathbb{E}\big[\tau^x(X) \mid \phi\big] = \tau^\phi(\phi), \tag{6}$$

so this stage directly learns the target estimand introduced in (4). In other words, the model learns how treatment benefit varies with the information in the multimodal representations.

Stage 3: Inference-Time Treatment Assignment. At inference time, only the multimodal representation ϕ is required. Each EHR is first encoded into its representation, after which the trained second-stage model predicts the corresponding coarsened treatment effect. Treatment is then assigned to the top-k individuals with the highest predicted benefit.

5. Experiments

We describe the experimental setup and then evaluate AACE on synthetic, semi-synthetic, and real-world multimodal EHR benchmarks. Full details are provided in Appendices C–E.[1]

5.1. Datasets

SynSum. The fully synthetic SynSum dataset consists of 10,000 medical records with tabular variables and free-text clinical notes (Rabaey et al., 2024). The notes are generated with an LLM and describe patient symptoms such as *fever*, *cough*, and *pain*, which act as text-based confounders and influence both treatment and outcome. We generate the treatments and outcomes such that confounding is substantial and baseline risk is only moderately aligned with treatment benefit. Expert annotations are simulated using the true symptom values used to generate the text, and we additionally study how performance changes when these annotations are imperfect. This results in a controlled multimodal EHR benchmark with known ground-truth treatment effects.

MIMIC-Syn. The semi-synthetic MIMIC-Syn benchmark is derived from the publicly available MIMIC-III database of intensive care EHRs (Johnson et al., 2016). Following Chen et al. (2024), we retain the real clinical notes and demographic variables to construct the multimodal EHRs. We then generate treatments and outcomes using cardiovascular diagnoses, age, and sex. The diagnosis indicators act as text-based confounders and influence both treatment and outcome, while being only implicitly expressed in the clinical text. Expert annotations are simulated using these diagnosis indicators. This results in a semi-synthetic multimodal EHR benchmark with known ground-truth treatment effects and realistic covariate distributions.

1. Code available at https://github.com/henriarnoUG/causal-ehr.

MIMIC-Real. The MIMIC-Real dataset was also derived from MIMIC-III (Johnson et al., 2016). Each multimodal EHR consists of demographic variables and clinical notes recorded prior to treatment. The treatment corresponds to *antibiotic administration* during ICU admission, and the outcome is *in-hospital mortality*. Vital-sign measurements are used as additional training-time covariates, as they plausibly influence both treatment and outcomes. Since this is an observational dataset, the required unconfoundedness assumption cannot be verified, and ground-truth effects are unknown. We therefore **cannot** use MIMIC-Real to evaluate which policy is best, and instead use it *only* to qualitatively compare the treatment assignments of the different methods in a real-world setting. The goal of this experiment is therefore to study **how treatment decisions differ** in practice, rather than to determine which policy is optimal.

5.2. Baselines and Evaluation

Baselines. We compare AACE to two standard strategies for treatment policy learning from multimodal EHR data. First, we consider *risk-based models*, which prioritize patients based on predicted baseline risk. Second, we consider *representation-based causal models*, which estimate treatment effects directly from the multimodal representations, implicitly assuming that these representations are sufficient for confounding adjustment. For this class, we apply standard neural CATE estimators, namely TARNet (Shalit et al., 2017), DragonNet (Shi et al., 2019), and DR-learner (Kennedy, 2023), directly to the multimodal representations. All methods use the same frozen ModernBERT-based multimodal representations as input for fair comparison.

Evaluation. For datasets with known ground-truth effects (SynSum and MIMIC-Syn), we evaluate the treatment policies induced by each method under varying treatment budgets. Each method produces a ranking of individuals, either by predicted treatment benefit or by baseline risk, and we evaluate the policy obtained by treating the top-k individuals. We report four metrics: AUTOC (area under the treatment-outcome curve) (Yadlowsky et al., 2025), which measures ranking quality across treatment budgets; the mean policy value across all budgets $(\overline{V})$; the policy value with a treatment budget of 10% of the population ($V_{10\%}$); and PEHE (precision in estimating heterogeneous effects), which measures treatment effect estimation error. Lower values for all metrics indicate better performance.

5.3. Results on SynSum

The SynSum benchmark provides a controlled setting with substantial confounding and moderate alignment between baseline risk and treatment benefit. Table 1 reports the policy evaluation results across training sizes. AACE consistently outperforms both risk-based and representation-based causal baselines on all metrics and at all training sizes. In particular, methods that estimate treatment effects directly from the multimodal representations (DR-Emb, DragonNet, TARNet) underperform compared to AACE, indicating that the additional *targeted* supervision is beneficial for confounding adjustment on this benchmark. Additional results in Appendix D show that the confounders are partially recoverable from the text representations, and report the policy metrics for additional training sizes.

Table 1: **SynSum (main results).** Policy metrics across training sizes, evaluated on a fixed test set (mean ± std. dev. over five random seeds). Lower values are better. PEHE measures treatment effect estimation error and is not defined for the risk-based model. The best model within each training size is shown in **bold**. The *oracle* row reports the metrics of a policy that ranks by the true treatment effect.

Training size	Model	AUTOC (↓)	$\overline{V}$ (↓)	$V_{10\%}$ (↓)	PEHE (↓)
1,000	Risk	-0.25 ± 0.02	0.77 ± 0.01	1.17 ± 0.01	—
	DR-Emb	-0.42 ± 0.04	0.69 ± 0.01	1.14 ± 0.01	0.71 ± 0.02
	DragonNet	-0.09 ± 0.16	0.79 ± 0.04	1.21 ± 0.04	0.96 ± 0.02
	TARNet	-0.43 ± 0.03	0.69 ± 0.01	1.14 ± 0.01	0.70 ± 0.03
	AACE (ours)	$\mathbf{-0.49 \pm 0.02}$	$\mathbf{0.68 \pm 0.01}$	$\mathbf{1.12 \pm 0.01}$	$\mathbf{0.62 \pm 0.02}$
4,000	Risk	-0.22 ± 0.01	0.77 ± 0.00	1.18 ± 0.00	—
	DR-Emb	-0.51 ± 0.01	0.66 ± 0.00	1.12 ± 0.00	0.59 ± 0.02
	DragonNet	-0.40 ± 0.02	0.70 ± 0.00	1.14 ± 0.01	0.71 ± 0.02
	TARNet	-0.51 ± 0.02	0.66 ± 0.00	1.13 ± 0.01	0.58 ± 0.02
	AACE (ours)	$\mathbf{-0.56 \pm 0.01}$	$\mathbf{0.65 \pm 0.00}$	$\mathbf{1.11 \pm 0.01}$	$\mathbf{0.53 \pm 0.02}$
8,000	Risk	-0.23 ± 0.00	0.77 ± 0.00	1.18 ± 0.01	—
	DR-Emb	-0.53 ± 0.01	0.65 ± 0.00	1.12 ± 0.00	0.56 ± 0.01
	DragonNet	-0.48 ± 0.01	0.67 ± 0.00	1.13 ± 0.00	0.63 ± 0.02
	TARNet	-0.53 ± 0.00	0.65 ± 0.00	1.12 ± 0.00	0.58 ± 0.02
	AACE (ours)	$\mathbf{-0.58 \pm 0.00}$	$\mathbf{0.64 \pm 0.00}$	$\mathbf{1.11 \pm 0.00}$	$\mathbf{0.50 \pm 0.01}$
Oracle		−0.74	0.61	1.08	0.00

5.4. Results on MIMIC-Syn

Table 2 reports the results on the semi-synthetic MIMIC-Syn benchmark. AACE again achieves the strongest performance overall. It obtains the best AUTOC, $V_{10\%}$, and PEHE, while remaining highly competitive on the mean policy value, where DragonNet performs marginally better. These results show that the advantage of annotation-assisted learning is not limited to the synthetic setting, but persists in a more realistic benchmark with real multimodal covariate distributions. Additional results in Appendix D show that the confounders are partially recoverable from the clinical text representations.

Table 2: **MIMIC-Syn (main results).** Policy metrics on a fixed test set (mean ± std. dev. over twenty random seeds). See Table 1 for metric definitions and interpretation.

Model	AUTOC (↓)	$\overline{V}$ (↓)	$V_{10\%}$ (↓)	PEHE (↓)
Risk	0.03 ± 0.00	6.49 ± 0.00	6.96 ± 0.00	—
DR-Emb	-0.12 ± 0.06	6.45 ± 0.01	6.92 ± 0.02	0.60 ± 0.04
DragonNet	-0.16 ± 0.04	$\mathbf{6.44 \pm 0.01}$	6.91 ± 0.01	0.57 ± 0.04
TARNet	-0.16 ± 0.04	6.44 ± 0.01	6.91 ± 0.01	0.60 ± 0.06
AACE (ours)	$\mathbf{-0.17 \pm 0.06}$	6.45 ± 0.01	$\mathbf{6.90 \pm 0.02}$	$\mathbf{0.45 \pm 0.01}$
Oracle	−0.43	6.37	6.84	0.00

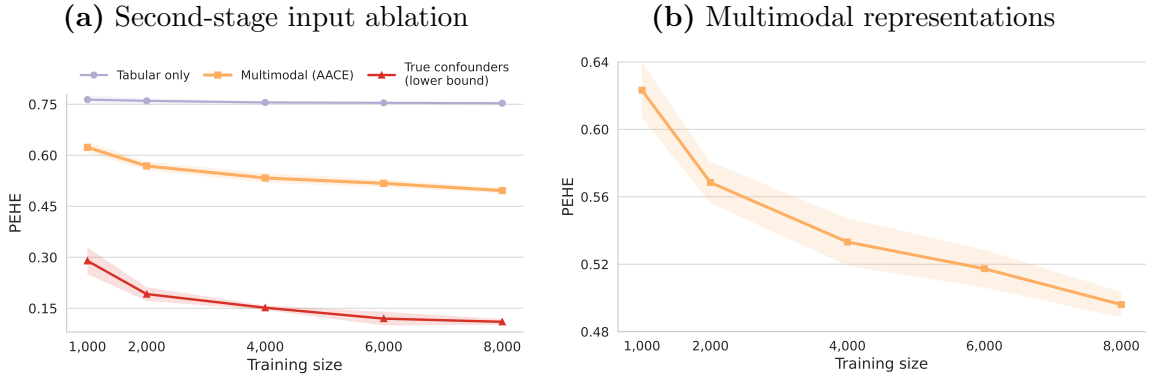

Figure 3: **SynSum (learning coarsened treatment effects).** PEHE as a function of training size. Panel (a) reports a second-stage input ablation in which the same DR pseudo-outcomes are used throughout, while the effect model is trained using tabular variables only, multimodal representations, or the true confounders, which are not available in practice and therefore provide a lower bound on PEHE. Panel (b) shows a zoomed view of the multimodal setting used by AACE, where PEHE decreases steadily with training size. Lower values are better.

5.5. Robustness Analyses

We further analyze how AACE depends on the quality of its annotations and multimodal representations. All analyses in this subsection are performed on the SynSum dataset, extended results are reported in Appendix D.

Input Ablation. Figure 3 reports a second-stage input ablation in which the same DR pseudo-outcomes are used throughout, while the effect model is trained using either (i) only the tabular variables from the EHR, (ii) the multimodal representations used by AACE, or (iii) the true confounders, which are not available in practice and therefore provide a lower bound on PEHE. The multimodal representations substantially reduce treatment effect estimation error relative to tabular-only inputs, and the error decreases steadily as more training data becomes available. This suggests that the multimodal representations capture relevant confounding signal for learning the coarsened treatment effects.

Annotation Quality. We next assess the robustness of AACE to imperfect annotations. We consider three failure modes that may arise in practice: (i) noisy or incorrect annotations of relevant confounders, simulated by replacing annotations with random values drawn from the empirical covariate distribution, (ii) missing relevant confounders, simulated by removing annotated confounders before training, and (iii) irrelevant annotations, simulated by adding variables to the annotations that are unrelated to the true confounding structure. Tables 3–5 show that noisy annotations and missing relevant confounders lead to gradual performance degradation, whereas adding irrelevant variables has little effect. Overall, these results indicate that AACE mainly depends on capturing the relevant confounding signal, while being relatively insensitive to irrelevant annotations.

Table 3: **SynSum (noisy annotations).** AUTOC of AACE as a function of annotation corruption, using 4,000 training samples and a fixed test set (mean ± std. dev. over five random seeds). Lower values are better.

Corr. (%)	0	5	10	25	50	100
AUTOC	-0.56 ± 0.01	-0.55 ± 0.01	-0.53 ± 0.01	-0.49 ± 0.01	-0.42 ± 0.04	-0.37 ± 0.06

Table 4: **SynSum (missing relevant confounders).** AUTOC of AACE as a function of the number of removed annotated confounders. See Table 3 for details.

Dropped (n)	0	1	2	3	4	5
AUTOC	-0.56 ± 0.01	-0.51 ± 0.03	-0.50 ± 0.01	-0.47 ± 0.04	-0.41 ± 0.03	-0.38 ± 0.05

Table 5: **SynSum (irrelevant annotations).** AUTOC of AACE as a function of the number of added irrelevant annotation variables. See Table 3 for details.

Added (n)	0	1	2	3	4	5
AUTOC	-0.56 ± 0.01	-0.56 ± 0.01	-0.56 ± 0.01	-0.56 ± 0.01	-0.56 ± 0.01	-0.55 ± 0.01

Representation Quality. We next assess robustness to the quality and choice of the multimodal representations. First, we degrade representation quality by randomly masking a fraction of the embedding dimensions. Second, we replace the original ModernBERT embeddings with Gemma embeddings (Vera et al., 2025). Tables 6 and 7 show that (i) AACE degrades gradually under representation masking, and (ii) the relative ordering of methods remains largely unchanged with Gemma embeddings. This suggests that the benefits of AACE are not specific to the original encoder, but depend on the representations retaining relevant confounding signal.

Table 6: **SynSum (representation masking).** AUTOC of AACE as a function of the fraction of embedding dimensions randomly masked. See Table 3 for details.

Masking (%)	0	75	90	95	100
AUTOC	-0.56 ± 0.01	-0.54 ± 0.01	-0.51 ± 0.01	-0.46 ± 0.00	-0.32 ± 0.01

Table 7: **SynSum (alternative encoder).** AUTOC of all methods using ModernBERT and Gemma embeddings. See Table 3 for details.

Encoder	Risk	DR-Emb	DragonNet	TARNet	AACE
ModernBERT	-0.22 ± 0.01	-0.51 ± 0.01	-0.40 ± 0.02	-0.51 ± 0.02	$\mathbf{-0.56 \pm 0.01}$
Gemma	-0.25 ± 0.01	-0.57 ± 0.01	-0.48 ± 0.00	-0.54 ± 0.03	$\mathbf{-0.60 \pm 0.00}$

Table 8: **MIMIC-Real (policy overlap).** Policy overlap and ranking agreement on real-world data (mean ± std. dev. over five random seeds). Jaccard measures agreement between top-k treatment assignments, while Spearman ρ captures global rank correlation.

Metric	AACE vs TARNet	AACE vs Risk
Jaccard overlap		
$k = 5\%$	0.088 ± 0.106	0.037 ± 0.027
$k = 10\%$	0.155 ± 0.115	0.076 ± 0.048
$k = 20\%$	0.245 ± 0.088	0.141 ± 0.052
$k = 50\%$	0.573 ± 0.075	0.362 ± 0.051
Spearman correlation		
ρ	0.551 ± 0.202	0.067 ± 0.170

5.6. Results on MIMIC-Real

Table 8 reports agreement between the treatment rankings induced by AACE, TARNet, and the risk-based model on the MIMIC-Real dataset. As discussed above, these results do **not** allow us to determine which policy is best, since ground-truth treatment effects are unknown. Instead, they show that *the different modeling strategies induce materially different treatment assignments in practice*. AACE is more strongly aligned with the representation-based causal baseline than with the risk-based model, indicating that causal and risk-based strategies prioritize substantially different patient groups in practice. Additional comparisons, including agreement with observed treatment assignments and the full pairwise analysis, are reported in Appendix D.

6. Discussion

Empirical Findings. Across synthetic and semi-synthetic benchmarks, AACE outperforms both risk-based and representation-based causal baselines, and remains robust to moderately imperfect annotations and representations. On real-world EHR data, the different methods induce substantially different treatment assignments, suggesting that the choice of modeling strategy can directly affect which patients are prioritized for treatment. Together, these results show that targeted training-time supervision can substantially improve treatment policy learning in multimodal settings where representations alone are insufficient for confounding adjustment.

Technical Implications. We study an important yet understudied setting in causal inference: multimodal data, where learned representations may not be sufficient for confounding adjustment. In that case, treatment effects are *not* identified from the observed data. Applying causal estimators directly to the representations may still perform well empirically, but the resulting treatment effect estimates remain *biased*, and the magnitude of this bias is generally unknown. In this setting, confounding must be addressed using additional training-time covariates that are sufficient for confounding adjustment. We argue that expert annotations offer a practical way to obtain such covariates.

Clinical Relevance. Electronic health records are a natural choice for learning data-driven treatment policies, because they are widely available and contain rich patient-specific information. However, learning *effective* and *reliable* treatment policies requires proper confounding adjustment, which is non-trivial in multimodal EHRs where relevant confounders may be distributed across tabular data and text. With AACE, we argue that expert knowledge remains important in this setting. By using expert-provided annotations during training, AACE offers a practical way to support confounding adjustment while still operating on standard multimodal EHR inputs at inference. This makes the approach compatible with standard clinical workflows, since the expert input is required only during model development, while routine EHR data are used for deployment. As with any clinical decision-support model, applying AACE in a new setting requires careful evaluation of the learned policy in the target population and clinical context.

Limitations and Future Work. Our study has several limitations. First, as noted in Section 3 and illustrated in Appendix A, we do not propose a universal annotation protocol for multimodal EHRs. This is deliberate, since the required training-time covariates for confounding adjustment depend on the healthcare application studied. Which covariates require annotation in practice should be guided by domain expertise. Second, as is standard in causal inference, our quantitative evaluation is limited to synthetic and semi-synthetic benchmarks where ground-truth treatment effects are available, while the analysis on real-world data remains qualitative. A useful direction for future work is therefore to evaluate AACE on RCT data, where policy performance can be measured from observed outcomes. Third, our experiments rely on simulated annotations or proxy training-time covariates rather than expert-provided annotations. An important next step is therefore to evaluate the method with real annotations, as well as more scalable forms of supervision such as partial annotation or LLM-assisted labeling with expert verification.

7. Conclusion

In this work, we study treatment policy learning from multimodal electronic health records (EHRs), where confounding information is distributed across tabular data and clinical text. We proposed AACE, an annotation-assisted method that uses expert-provided annotations during training to support confounding adjustment, while relying only on multimodal representations at inference time. Across synthetic and semi-synthetic benchmarks, AACE outperforms both risk-based and representation-based causal baselines, while remaining robust to moderately imperfect annotations and representations. On real-world data, we find that the choice of modeling strategy substantially affects the resulting treatment assignments, highlighting that method choice matters in practice. Together, these results suggest that incorporating expert knowledge during training is a promising direction for improving treatment policy learning from multimodal healthcare data.

Acknowledgments

This work received funding from the Research Foundation Flanders (FWO Vlaanderen) with grant number 11Q2C24N and from the Flemish government under the “Onderzoeksprogramma Artificiele Intelligentie (AI) Vlaanderen” program.

References

Ahmed Alaa and Mihaela van der Schaar. Bayesian inference of individualized treatment effects using multi-task gaussian processes. In *Advances in Neural Information Processing Systems*, 2017.

Henri Arno, Dennis Frauen, Emil Javurek, Thomas Demeester, and Stefan Feuerriegel. Rank-learner: Orthogonal ranking of treatment effects. In *The International Conference on Machine Learning*, 2026.

Susan Athey and Stefan Wager. Policy learning with observational data. *Econometrica*, 89 (1), 2021.

Susan Athey, Niall Keleher, and Jann Spiess. Machine learning who to nudge: Causal vs predictive targeting in a field experiment on student financial aid renewal. *Journal of Econometrics*, 249, 2025.

Ioana Bica, Ahmed Alaa, Craig Lambert, and Mihaela Van Der Schaar. From real-world patient data to individualized treatment effects using machine learning: current and future methods to address underlying challenges. *Clinical Pharmacology & Therapeutics*, 109(1), 2021.

Jacob Chen, Rohit Bhattacharya, and Katherine Keith. Proximal causal inference with text data. In *Advances in Neural Information Processing Systems*, 2024.

Alicia Curth and Mihaela van der Schaar. On inductive biases for heterogeneous treatment effect estimation. In *Advances in Neural Information Processing Systems*, 2021.

Adel Daoud, Connor Jerzak, and Richard Johansson. Conceptualizing treatment leakage in text-based causal inference. In *The North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2022.

Naoki Egami, Christian Fong, Justin Grimmer, Margaret Roberts, and Brandon Stewart. How to make causal inferences using texts. *Science Advances*, 8(42), 2022.

Andre Esteva, Alexandre Robicquet, Bharath Ramsundar, Volodymyr Kuleshov, Mark Depristo, Katherine Chou, Claire Cui, Greg Corrado, Sebastian Thrun, and Jeff Dean. A guide to deep learning in healthcare. *Nature medicine*, 25(1), 2019.

Carlos Fernández-Loría and Foster Provost. Causal classification: Treatment effect estimation vs. outcome prediction. *Journal of Machine Learning Research*, 23(59), 2022.

Carlos Fernández-Loría, Foster Provost, Jesse Anderton, Benjamin Carterette, and Praveen Chandar. A comparison of methods for treatment assignment with an application to playlist generation. *Information Systems Research*, 34(2), 2023.

Stefan Feuerriegel, Dennis Frauen, Valentyn Melnychuk, Jonas Schweisthal, Konstantin Hess, Alicia Curth, Stefan Bauer, Niki Kilbertus, Isaac Kohane, and Mihaela van der Schaar. Causal machine learning for predicting treatment outcomes. *Nature Medicine*, 30(4), 2024.

Dennis Frauen, Valentyn Melnychuk, Jonas Schweisthal, Mihaela van der Schaar, and Stefan Feuerriegel. Treatment effect estimation for optimal decision-making. In *Advances in Neural Information Processing Systems*, 2025.

Benjamin Goldstein, Ann Marie Navar, Michael Pencina, and John Ioannidis. Opportunities and challenges in developing risk prediction models with electronic health records data: a systematic review. *Journal of the American Medical Informatics Association*, 24(1), 2016.

Limor Gultchin, David Watson, Matt Kusner, and Ricardo Silva. Operationalizing complex causes: A pragmatic view of mediation. In *The International Conference on Machine Learning*, 2021.

Connor Jerzak, Fredrik Johansson, and Adel Daoud. Image-based treatment effect heterogeneity. In *The Conference on Causal Learning and Reasoning*, 2023.

Fredrik Johansson, Uri Shalit, and David Sontag. Learning representations for counterfactual inference. In *The International Conference on Machine Learning*, 2016.

Alistair Johnson, Tom Pollard, Lu Shen, Li-wei Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger Mark. MIMIC-III, a freely accessible critical care database. *Scientific Data*, 3(160035), 2016.

Nathan Kallus. Balanced policy evaluation and learning. In *Advances in Neural Information Processing Systems*, 2018.

Katherine Keith, David Jensen, and Brendan O'Connor. Text and causal inference: A review of using text to remove confounding from causal estimates. In *The Annual Meeting of the Association for Computational Linguistics*, 2020.

Edward Kennedy. Towards optimal doubly robust estimation of heterogeneous causal effects. *Electronic Journal of Statistics*, 17(2), 2023.

Sven Klaassen, Jan Teichert-Kluge, Philipp Bach, Victor Chernozhukov, Martin Spindler, and Suhas Vijaykumar. DoubleMLdeep: Estimation of causal effects with multimodal data, 2024. preprint - arXiv:2402.01785.

Sören Künzel, Jasjeet Sekhon, Peter Bickel, and Bin Yu. Meta learners for estimating heterogeneous treatment effects using machine learning. *Proceedings of the National Academy of Sciences*, 116(10), 2019.

Jingshu Liu, Zachariah Zhang, and Narges Razavian. Deep EHR: Chronic disease prediction using medical notes. In *Machine Learning for Healthcare Conference*, 2018.

David Lopez-Paz, Léon Bottou, Bernhard Schölkopf, and Vladimir Vapnik. Unifying distillation and privileged information. In *The International Conference on Learning Representations*, 2016.

Yuchen Ma, Dennis Frauen, Jonas Schweisthal, and Stefan Feuerriegel. LLM-driven treatment effect estimation under inference time text confounding. In *Advances in Neural Information Processing Systems*, 2025.

Maggie Makar, Adith Swaminathan, and Emre Kıcıman. A distillation approach to data efficient individual treatment effect estimation. In *The AAAI Conference on Artificial Intelligence*, 2019.

Valentyn Melnychuk, Dennis Frauen, and Stefan Feuerriegel. Bounds on representation-induced confounding bias for treatment effect estimation. In *The International Conference on Learning Representations*, 2024.

Xinkun Nie and Stefan Wager. Quasi-oracle estimation of heterogeneous treatment effects. *Biometrika*, 108(2), 2020.

Xinkun Nie, Emma Brunskill, and Stefan Wager. Learning when-to-treat policies. *Journal of the American Statistical Association*, 116(533), 2021.

Min Qian and Susan Murphy. Performance guarantees for individualized treatment rules. *Annals of statistics*, 39(2), 2011.

Paloma Rabaey, Henri Arno, Stefan Heytens, and Thomas Demeester. SynSum – Synthetic benchmark with structured and unstructured medical records. In *AAAI 2025 Workshop on GenAI4Health*, 2024.

Alvin Rajkomar, Jeffrey Dean, and Isaac Kohane. Machine learning in medicine. *New England Journal of Medicine*, 380(14), 2019.

Donald Rubin. Causal inference using potential outcomes. *Journal of the American Statistical Association*, 100(469), 2005.

Pedro Sanchez and Sotirios Tsaftaris. Diffusion causal models for counterfactual estimation. In *The Conference on Causal Learning and Reasoning*, 2022.

Uri Shalit, Fredrik Johansson, and David Sontag. Estimating individual treatment effect: generalization bounds and algorithms. In *The International Conference on Machine Learning*, 2017.

Claudia Shi, David Blei, and Victor Veitch. Adapting neural networks for the estimation of treatment effects. In *Advances in Neural Information Processing Systems*, 2019.

Benjamin Shickel, Patrick James Tighe, Azra Bihorac, and Parisa Rashidi. Deep EHR: A survey of recent advances in deep learning techniques for electronic health record (EHR) analysis. *IEEE journal of biomedical and health informatics*, 22(5), 2017.

Alice Tang, Sarah Woldemariam, Silvia Miramontes, Beau Norgeot, Tomiko Oskotsky, and Marina Sirota. Harnessing EHR data for health research. *Nature Medicine*, 30(7), 2024.

Maarten Van Smeden, Johannes Reitsma, Richard Riley, Gary Collins, and Karel Moons. Clinical prediction models: diagnosis versus prognosis. *Journal of clinical epidemiology*, 132, 2021.

Victor Veitch, Dhanya Sridhar, and David Blei. Adapting text embeddings for causal inference. In *Uncertainty in Artificial Intelligence*, 2020.

Henrique Schechter Vera, Sahil Dua, Biao Zhang, Daniel Salz, Ryan Mullins, Sindhu Raghuram Panyam, Sara Smoot, Iftekhar Naim, Joe Zou, Feiyang Chen, et al. EmbeddingGemma: Powerful and lightweight text representations, 2025. preprint - arXiv:2509.20354.

Stefan Wager and Susan Athey. Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 113(523), 2018.

Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, Nathan Cooper, Griffin Adams, Jeremy Howard, and Iacopo Poli. Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference, 2024. preprint - arXiv:2412.13663.

Zach Wood-Doughty, Ilya Shpitser, and Mark Dredze. Challenges of using text classifiers for causal inference. In *Empirical Methods in Natural Language Processing*, 2018.

Steve Yadlowsky, Scott Fleming, Nigam Shah, Emma Brunskill, and Stefan Wager. Evaluating treatment prioritization rules via rank-weighted average treatment effects. *Journal of the American Statistical Association*, 120(549), 2025.

Bo Yang and Lijun Wu. How to leverage the multimodal EHR data for better medical prediction. In *Empirical Methods in Natural Language Processing*, 2021.

Jinsung Yoon, James Jordon, and Mihaela van der Schaar. GANITE: Estimation of individualized treatment effects using generative adversarial nets. In *The International Conference on Learning Representations*, 2018.

Fucheng Warren Zhu, Connor Jerzak, and Adel Daoud. Optimizing multi-scale representations to detect effect heterogeneity using earth observation and computer vision: Application to two anti-poverty rcts. In *The Conference on Causal Learning and Reasoning*, 2025.

Appendix A. Annotation Procedure

As discussed in Section 3, when clinically relevant confounders are expressed in text but not reliably preserved in the learned representation ϕ, treatment effects are not identified from observational data (ϕ, T, Y) alone. Treatment effect estimation in this multimodal setting therefore requires training-time covariates that are sufficient for confounding adjustment. With AACE, we argue that expert-provided annotations offer a pragmatic way to obtain such covariates. Because the specific variables that need to be annotated depend on the clinical application and require domain expertise, we deliberately do not prescribe a fixed annotation procedure. Instead, we provide below an illustrative example of how such annotations may be obtained in one of our datasets.

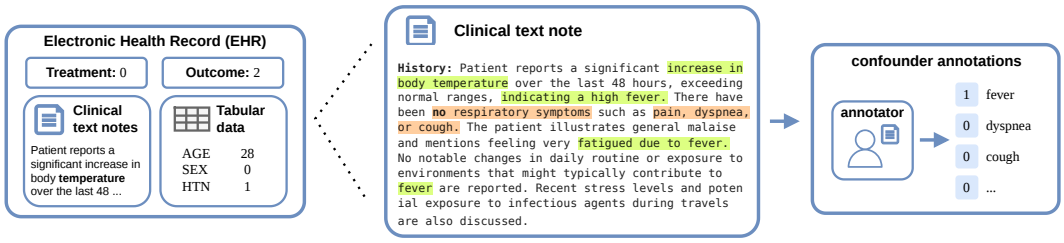

Figure 4: **SynSum (example annotation procedure).** Illustration of how clinically relevant confounding information may be expressed implicitly in a clinical note and annotated for use in AACE. In this example, highlighted spans provide evidence for symptom variables such as *fever*, *dyspnea*, and *cough*, which are used as training-time covariates for confounding adjustment.

Problem Setting. SynSum is a dataset that simulates primary-care encounters for patients with respiratory diseases (discussed in detail in Appendix C). Each simulated record contains both tabular variables and a free-text clinical note: background information, such as demographics, is recorded in structured form, while symptom variables such as *fever*, *dyspnea*, and *cough* are expressed in the note. In this setting, the treatment decision under study is antibiotics prescription, which is influenced by the symptoms that also affect patient outcomes, and therefore requires adjustment for them during effect estimation.

Example Annotation. Figure 4 illustrates one plausible *retrospective* annotation workflow for AACE, in which a domain expert identifies clinical factors from the note that may influence both antibiotic prescription and patient outcomes, and records them in structured form. In SynSum, this corresponds to the symptom variables. The highlighted spans in Figure 4 indicate the textual evidence supporting these variables, while the binary indicators illustrate how they may be represented as additional training-time covariates. In practice, the same variables could also be recorded directly as part of routine clinical documentation during *a one-time annotation effort*. In that case, the same domain expertise used to assess the patient and make the treatment decision could also be used to record a small set of structured covariates that are relevant for confounding adjustment. These covariates are then used, together with the tabular EHR variables, for confounding adjustment during training. At inference time, the annotations are not required and the policy depends only on the representation ϕ.

Appendix B. Proof

In this appendix, we provide the proof of Proposition 1.

Proof Assume that the true treatment effects have no ties. Let $x_{(1)}, \ldots, x_{(N)}$ denote the individuals ordered by decreasing treatment effects, so that

$$\tau^x(x_{(1)}) > \tau^x(x_{(2)}) > \cdots > \tau^x(x_{(N)}). \tag{7}$$

For a fixed treatment budget $k < N$, define the treatment effect margin at the decision threshold as

$$\gamma_k = \tau^x(x_{(k)}) - \tau^x(x_{(k+1)}), \tag{8}$$

and let

$$\delta_i = \left|\tau^x(x_i) - \tau^\phi(\phi_i)\right| \tag{9}$$

denote the coarsening bias for individual i.

Let $S_k = \{x_{(1)}, \ldots, x_{(k)}\}$ denote the true top-k according to $\tau^x(x)$. To prove the proposition, it suffices to show that every individual in S_k remains ranked above every individual outside S_k when ranking by the coarsened effects $\tau^\phi(\phi)$. Now consider any $x_i \in S_k$ and any $x_j \notin S_k$. By construction we have,

$$\tau^x(x_i) \geq \tau^x(x_{(k)}) \qquad \text{and} \qquad \tau^x(x_j) \leq \tau^x(x_{(k+1)}). \tag{10}$$

Therefore,

$$\begin{aligned} \tau^x(x_i) - \tau^x(x_j) &\geq \tau^x(x_{(k)}) - \tau^x(x_{(k+1)}) && (11)\\ &\geq \gamma_k. && (12) \end{aligned}$$

By definition of the coarsening bias,

$$\tau^\phi(\phi_i) \geq \tau^x(x_i) - \delta_i \qquad \text{and} \qquad \tau^\phi(\phi_j) \leq \tau^x(x_j) + \delta_j. \tag{13}$$

Subtracting yields

$$\begin{aligned} \tau^\phi(\phi_i) - \tau^\phi(\phi_j) &\geq \big(\tau^x(x_i) - \delta_i\big) - \big(\tau^x(x_j) + \delta_j\big) && (14)\\ &= \big(\tau^x(x_i) - \tau^x(x_j)\big) - (\delta_i + \delta_j) && (15)\\ &\geq \gamma_k - 2\max_m \delta_m. && (16) \end{aligned}$$

By the assumption of the proposition,

$$2\max_m \delta_m < \gamma_k, \tag{17}$$

and therefore

$$\tau^\phi(\phi_i) - \tau^\phi(\phi_j) > 0. \tag{18}$$

Hence,

$$\tau^\phi(\phi_i) > \tau^\phi(\phi_j). \tag{19}$$

Since this holds for every $x_i \in S_k$ and every $x_j \notin S_k$, every true top-k individual remains ranked above every individual outside the top-k under the coarsened effects. Therefore, ranking by $\tau^\phi(\phi)$ yields exactly the same top-k treatment set as ranking by $\tau^x(x)$. ■

Appendix C. Datasets

This appendix provides a detailed description of the datasets used in our experiments: the fully synthetic SynSum dataset (Rabaey et al., 2024), the semi-synthetic MIMIC-Syn dataset (Johnson et al., 2016; Chen et al., 2024), and the real-world MIMIC-Real dataset (Johnson et al., 2016). For the synthetic and semi-synthetic datasets, we describe the data-generating process, including the construction of treatments and outcomes. The MIMIC-Real dataset is included **only** for qualitative comparison of the treatment assignment strategies, since ground-truth treatment effects are unknown.

C.1. SynSum

The fully synthetic SynSum dataset (Rabaey et al., 2024)[2] consists of 10,000 medical patient records simulating primary care encounters for patients with respiratory diseases. Each record combines tabular variables with an associated free-text clinical note, mimicking the multimodal structure of real electronic health records (EHRs). The dataset is entirely specified and reproducible: tabular variables are sampled from a Bayesian network defined by a domain expert, while clinical notes are generated using GPT-4. SynSum thus provides realistic, simulated EHR data with ground-truth counterfactuals, enabling controlled evaluation of the considered treatment assignment strategies.

Tabular Variables. In SynSum, the structured variables are sampled from a Bayesian network that defines the underlying clinical dependencies in the simulated population. The variables include diagnoses (pneumonia and common cold), symptoms (dyspnea, cough, pain, fever, and nasal congestion), underlying respiratory conditions (asthma, smoking, chronic obstructive pulmonary disease, and hay fever), and non-clinical factors (season, prescription policy, and self-employment). Each variable is binary and represented by an indicator $x_{\text{var}} \in \{0, 1\}$. Full details of the Bayesian network and its parametrization are given in Rabaey et al. (2024). The structured variables play different roles in the benchmark. A subset of them (self-employment status, smoking, winter season, asthma, chronic obstructive pulmonary disease, and hay fever) appears as tabular variables in the simulated EHRs, representing background and chronic-condition information typically recorded in structured form. The symptom variables (dyspnea, cough, pain, fever, and nasal congestion) are used to generate the accompanying clinical text, where they are expressed implicitly rather than as structured fields. Finally, a small number of variables (the diagnoses and the policy indicator) are used only to induce realistic dependencies in the data-generating process and are not included directly in the final EHR.

Treatment and Outcome. Based on the structured variables, we simulate treatment assignment and potential outcomes. Relative to the original SynSum data, we define a new treatment and outcome generating process tailored to treatment policy learning, with stronger confounding and weaker alignment between baseline risk and treatment benefit. The treatment variable $T \in \{0, 1\}$ indicates whether antibiotics are prescribed during the simulated consultation. It is sampled from a logistic regression model

$$P(T = 1 \mid X = x) = \sigma\big(\ell(x)\big), \tag{20}$$

2. https://github.com/prabaey/SimSum

where $\sigma(\cdot)$ denotes the sigmoid function and the logit is given by

$$\begin{aligned}\ell(x) = &-2.2 + 2.0\, x_{\text{pol}} + 1.5\, x_{\text{dysp}} + 1.1\, x_{\text{cough}} + 1.3\, x_{\text{f\_high}} + 0.6\, x_{\text{pain}} \\ &+ 0.9\, x_{\text{pol}}\, x_{\text{dysp}} + 0.7\, x_{\text{pol}}\, x_{\text{f\_high}} - 0.5\, x_{\text{nasal}}\, x_{\text{f\_low}}.\end{aligned} \tag{21}$$

This specification reflects realistic clinical behaviour, where treatment is more likely for patients with more severe symptoms, especially under permissive prescription policies. We then define the untreated potential outcome through a baseline response surface $\mu_0(x)$,

$$\begin{aligned}\mu_0(x) = &\, 1.8\, x_{\text{self}} + 1.6\, x_{\text{dysp}} + 1.2\, x_{\text{cough}} + 0.5\, x_{\text{pain}} + 0.4\, x_{\text{nasal}} \\ &+ 0.3\, x_{\text{f\_low}} + 1.5\, x_{\text{f\_high}} + 1.1\, x_{\text{dysp}}\, x_{\text{f\_high}} \\ &+ 0.9\, x_{\text{self}}\, x_{\text{cough}} - 0.6\, x_{\text{nasal}}\, x_{\text{f\_low}} \\ &+ 0.7\, x_{\text{pain}}\, x_{\text{self}} + 0.8\, x_{\text{cough}}\, x_{\text{f\_high}} + 0.7\, x_{\text{dysp}}\, x_{\text{self}}.\end{aligned} \tag{22}$$

The conditional average treatment effect (CATE) is specified as

$$\tau^x(x) = -1 + 1.1\, x_{\text{pain}} + 1.0\, x_{\text{nasal}} + 0.7\, x_{\text{f\_low}} - 1.0\, x_{\text{dysp}} - 0.7\, x_{\text{cough}} - 1.1\, x_{\text{f\_high}}, \tag{23}$$

so that treatment is more beneficial for patients with more severe lower-respiratory symptoms, and less beneficial for those with milder or upper-airway symptoms. The treated response surface is then defined as

$$\mu_1(x) = \mu_0(x) + \tau^x(x). \tag{24}$$

Finally, the potential outcomes are generated by adding Gaussian noise,

$$Y(0) = \mu_0(x) + \varepsilon_0, \qquad Y(1) = \mu_1(x) + \varepsilon_1, \tag{25}$$

where $\varepsilon_0, \varepsilon_1 \sim \mathcal{N}(0, \sigma_y^2)$, and the observed outcome is

$$Y = T\, Y(1) + (1 - T)\, Y(0). \tag{26}$$

This construction ensures that symptom variables act as confounders by influencing both treatment assignment and outcome, while treatment benefit is heterogeneous and only partially aligned with baseline risk.

Text Generation. Each record's clinical note is generated with GPT-4 so that the patient's symptom profile is reflected in natural language. The prompts are designed to encourage clinical realism, coherence, and stylistic variation, resembling documentation written at the time of consultation. In the final EHR, these symptoms are not included as structured variables, but are expressed implicitly in the free-text note. Each simulated EHR therefore consists of tabular background variables (e.g., smoking status, chronic conditions, and season) together with a clinical note describing the patient's symptoms.

Feature Representations. Each simulated EHR in SynSum consists of tabular background variables and a clinical note. We encode the text using the pretrained language model ModernBERT (Warner et al., 2024), mean-pool the token embeddings, and concatenate the resulting text representation with the tabular variables to obtain the multimodal representation ϕ. This representation combines the structured background information with confounding signal expressed implicitly in the text.

Confounding Annotations. AACE assumes that domain experts can provide annotations of confounding information expressed in the clinical text. In SynSum, this process is simulated using the ground-truth symptom variables (e.g., fever, cough, and dyspnea), which influence both treatment assignment and outcome. This yields a controlled setting in which confounding adjustment is possible during training through the annotations, while inference relies only on the multimodal representations. In addition, we use this benchmark later to study how AACE degrades as annotation quality decreases.

C.2. MIMIC-Syn

The semi-synthetic MIMIC-Syn dataset is derived from the publicly available MIMIC-III database (Johnson et al., 2016)[3], which contains de-identified electronic health records from over 35,000 intensive care unit (ICU) patients. The dataset includes both structured variables and free-text clinical notes. Following Chen et al. (2024), we retain the real tabular and textual features while generating synthetic treatment and outcome variables. This design preserves the realism of real-world EHR data while enabling controlled evaluation under a known causal ground truth.

Data Preprocessing. We follow the preprocessing pipeline of Chen et al. (2024) to construct patient-level records from MIMIC-III. Starting from all clinical notes, we remove entries with missing hospital admission IDs and exclude discharge summaries. For each patient, identified by a unique subject ID, we retain only the earliest hospital admission and concatenate all associated clinical notes into a single text sequence. We additionally extract the demographic variables age and sex, which serve as the tabular covariates in the EHR. Patients younger than 18 or older than 100 years are excluded. Each resulting record therefore consists of a single aggregated clinical note and a small set of demographic variables, forming a multimodal EHR with real-world patient data. Other structured variables from MIMIC-III are used later in the data-generating process to construct treatment and outcome variables, but are not included directly in the final EHR.

Treatment and Outcome. To introduce heterogeneous treatment effects, we define a semi-synthetic treatment and outcome generating process on top of the real MIMIC-III covariates. Treatment assignment and potential outcomes are generated from four cardiovascular diagnoses (hypertension, coronary atherosclerosis, atrial fibrillation, and congestive heart failure) together with age and sex. The diagnosis indicators ($x_{\text{var}} \in \{0, 1\}$) are extracted from the ICD-9 codes in the MIMIC-III tables, but are not included in the constructed EHRs. Prior analyses by Chen et al. (2024) showed that these diagnoses are among the most predictable conditions from clinical text, indicating that they are implicitly expressed in the notes. Using these variables to generate treatments and outcomes therefore induces implicit multimodal confounding, analogous to the SynSum setting. The treatment variable $T \in \{0, 1\}$ indicates whether the hypothetical intervention is assigned. It is sampled from a logistic regression model

$$P(T = 1 \mid X = x) = \sigma\big(\ell(x)\big), \tag{27}$$

3. https://physionet.org/content/mimiciii/1.4/

where $\sigma(\cdot)$ denotes the sigmoid function and the logit is given by

$$\begin{aligned}\ell(x) = &-0.8 + 0.08\, x_{\text{age,c}} + 0.65\, x_{\text{sex}} + 0.50\, x_{\text{hyp}} + 0.60\, x_{\text{cor}} \\ &+ 0.40\, x_{\text{afib}} + 0.70\, x_{\text{chf}} + 0.25\, x_{\text{sex}}\, \mathbb{I}(x_{\text{age}} > 75) \\ &+ 0.25\, x_{\text{cor}}\, x_{\text{chf}},\end{aligned} \tag{28}$$

with $x_{\text{age,c}} = x_{\text{age}} - 65$. This specification reflects confounding by indication, where older patients and patients with greater cardiovascular burden are more likely to receive treatment, while also being at higher baseline risk. We define the untreated potential outcome through a baseline response surface $\mu_0(x)$,

$$\begin{aligned}\mu_0(x) = 5.5 &+ 0.10\, x_{\text{age,c}} + 0.60\, x_{\text{sex}} + 0.70\, x_{\text{hyp}} + 1.00\, x_{\text{cor}} + 0.85\, x_{\text{afib}} + 1.20\, x_{\text{chf}} \\ &+ 0.020\, \max(x_{\text{age}} - 75, 0)^2 + 0.50\, x_{\text{cor}}\, x_{\text{chf}} + 0.35\, x_{\text{afib}}\, x_{\text{chf}} \\ &+ 0.25\, x_{\text{hyp}}\, x_{\text{cor}} + 0.20\, x_{\text{sex}}\, x_{\text{cor}} + 0.015\, x_{\text{age,c}}\, x_{\text{afib}}.\end{aligned} \tag{29}$$

The conditional average treatment effect (CATE) is specified as

$$\begin{aligned}\tau^x(x) = &-1.0 - 0.35\, x_{\text{hyp}} - 0.50\, x_{\text{cor}} + 0.20\, x_{\text{afib}} + 0.45\, x_{\text{chf}} \\ &+ 0.015\, \max(x_{\text{age}} - 75, 0) + 0.05\, x_{\text{sex}} + 0.20\, x_{\text{afib}}\, x_{\text{chf}} - 0.35\, x_{\text{hyp}}\, x_{\text{cor}}.\end{aligned} \tag{30}$$

Negative values of $\tau^x(x)$ indicate treatment benefit, since lower outcomes are better. This specification makes treatment beneficial on average, while allowing the benefit to vary substantially across subgroups. The treated response surface is then defined as

$$\mu_1(x) = \mu_0(x) + \tau^x(x). \tag{31}$$

Finally, the potential outcomes are generated by adding Gaussian noise,

$$Y(0) = \mu_0(x) + \varepsilon_0, \qquad Y(1) = \mu_1(x) + \varepsilon_1, \tag{32}$$

where $\varepsilon_0, \varepsilon_1 \sim \mathcal{N}(0, \sigma_y^2)$, and the observed outcome is

$$Y = T\, Y(1) + (1 - T)\, Y(0). \tag{33}$$

This setup produces a semi-synthetic benchmark in which treatment assignment is confounded, treatment benefit is heterogeneous, and baseline risk is not sufficient to determine which patients benefit most from treatment.

Feature Representations. Each record in MIMIC-Syn consists of a clinical note and tabular demographic variables (age and sex). We encode the clinical text using the pretrained language model ModernBERT (Warner et al., 2024), mean-pool the token embeddings, and concatenate the resulting text representation with the tabular covariates to obtain the multimodal representation ϕ. This representation combines the demographic information with implicit confounding signal related to the latent cardiovascular diagnoses. Consequently, ϕ provides a coarsened view of the underlying factors influencing treatment and outcome, mirroring the partial observability typical of real-world EHR data.

Confounding Annotations. AACE assumes that domain experts can provide annotations of confounding information expressed in the clinical text. In MIMIC-Syn, this process is simulated using the ground-truth cardiovascular diagnoses (hypertension, coronary atherosclerosis, atrial fibrillation, and congestive heart failure), which influence both treatment assignment and outcome. This yields a controlled semi-synthetic setting in which confounding adjustment is possible during training through the annotations, while inference relies only on the multimodal representations ϕ.

C.3. MIMIC-Real

We construct a real-world observational dataset from the MIMIC-III database (Johnson et al., 2016) to examine whether annotation-assisted, risk-based, and representation-based causal treatment strategies diverge in practice. In this setting, both the tabular variables and the clinical text notes of the electronic health records (EHRs) are derived directly from real-world patient data. The treatment (*antibiotic administration*) and outcome (*in-hospital mortality*) are likewise **observed** rather than generated. This dataset complements the synthetic and semi-synthetic analyses by extending the comparison of risk-based and causal policies to fully real-world clinical data. Because ground-truth counterfactuals are unknown, **we cannot determine which policy is best** and restrict the analysis to a **qualitative comparison** of the treatment assignments induced by the different methods.

Data Preprocessing. We construct a patient-level dataset of EHRs from all intensive care unit (ICU) stays in MIMIC-III. Each record combines tabular variables and clinical notes describing the patient's state prior to treatment. We define a reference time t_0 three hours after ICU admission and a follow-up window Δt of six hours. All covariates are derived from data recorded up to t_0; treatment is defined within the interval $(t_0, t_0 + \Delta t)$; and the outcome is measured after $t_0 + \Delta t$. ICU stays that end before $t_0 + \Delta t$ are excluded. As tabular variables, we retain the demographic attributes age and sex, consistent with MIMIC-Syn. We exclude individuals younger than 18 years and retain only the first ICU stay of each patient. All clinical notes recorded before t_0 are extracted and concatenated into a single text sequence representing the documentation available at the time of the treatment decision. Each resulting EHR therefore consists of the tabular demographic variables age and sex together with an aggregated clinical note summarizing the patient's condition prior to treatment.

Treatment and Outcome. The treatment variable $T \in \{0, 1\}$ indicates whether at least one antibiotic prescription was administered during the interval $(t_0, t_0 + \Delta t)$. Antibiotic prescriptions are identified using a list of antibiotic drug names. The outcome variable $Y \in \{0, 1\}$ denotes in-hospital mortality occurring after $t_0 + \Delta t$, reflecting post-treatment mortality between the end of the treatment window and hospital discharge. In the resulting dataset (8,685 ICU stays), 18.5% of patients received antibiotics within the treatment window (T=1), and 12.0% died after the treatment window during the same hospitalization (Y=1). Mortality was 16.4% among treated patients and 11.0% among untreated patients. This higher mortality among treated patients does not imply a harmful treatment effect, but instead reflects *confounding*: antibiotics are typically administered to patients with suspected sepsis or severe infection, who are at higher risk of death. Treatment and outcome are therefore strongly associated through shared confounders such as abnormal vital signs.

Feature Representations. Each EHR in MIMIC-Real consists of the tabular demographic variables age and sex together with an aggregated clinical note describing the patient's condition prior to treatment. We encode the clinical text using the pretrained language model ModernBERT (Warner et al., 2024), mean-pool the token embeddings, and concatenate the resulting text representation with the tabular variables to obtain the multimodal representation ϕ. This representation combines explicit demographic information with implicit clinical signals present in the real-world EHR data.

Additional Training-Time Covariates. For AACE, we use pre-treatment vital signs as additional training-time covariates for confounding adjustment. These measurements plausibly influence both antibiotic administration and mortality: clinicians are more likely to prescribe antibiotics to patients presenting with abnormal physiological measurements, such as fever, tachycardia, or hypotension, and these same measurements are also associated with higher mortality risk. We therefore extract the following vital signs prior to treatment (before t_0): temperature, heart rate, respiratory rate, oxygen saturation, systolic blood pressure, and mean arterial pressure.

To assess whether this confounding information is implicitly expressed in the clinical text, we evaluate how well abnormal vital-sign patterns can be predicted from the text representations. Vital signs are discretized based on standard clinical thresholds: fever (temperature $\geq 37.5°$C), hypothermia (temperature $\leq 35.5°$C), tachycardia (heart rate ≥ 100 beats/min), tachypnea (respiratory rate ≥ 22 breaths/min), hypoxemia (oxygen saturation $< 92\%$), and hypotension (mean arterial pressure < 65 mmHg or systolic blood pressure < 90 mmHg). Each abnormality indicator is then predicted using only the ModernBERT text embeddings. The dataset is randomly divided into 80% training, 10% validation, and 10% test splits. As shown in Table 9, these physiological states are predictable from the clinical text, with all AUROC values exceeding 0.5. This indicates that clinically relevant confounding information is at least partly encoded in the notes. Accordingly, we use the continuous vital-sign measurements as additional training-time covariates for confounding adjustment during training of AACE. At inference time, the method relies only on the multimodal representations ϕ. We note, however, that this real-world dataset may still contain additional **unobserved confounders** beyond the available vital-sign measurements.

C.4. Summary

In summary, the three datasets (SynSum, MIMIC-Syn, and MIMIC-Real) enable evaluation of annotation-assisted, risk-based, and representation-based causal treatment assignment strategies across settings ranging from fully synthetic simulations to real-world clinical data. In the following section, we present additional experimental results and analyses based on these datasets.

Table 9: **MIMIC-Real (confounder predictability from text).** Prediction of physiological confounders from pre-t_0 clinical notes (mean ± std. over five random seeds). "Prev." denotes the positive-class prevalence in the test set. Each classifier's decision threshold was selected to maximize F1 score on the validation set.

Confounder	Prev.	F1	Acc.	AUPRC	AUROC
Fever	0.14	0.337 ± 0.007	0.607 ± 0.026	0.253 ± 0.003	0.680 ± 0.003
Hypothermia	0.04	0.026 ± 0.041	0.951 ± 0.013	0.097 ± 0.026	0.649 ± 0.024
Tachycardia	0.22	0.435 ± 0.006	0.617 ± 0.021	0.381 ± 0.002	0.679 ± 0.003
Tachypnea	0.26	0.467 ± 0.006	0.599 ± 0.025	0.408 ± 0.004	0.678 ± 0.005
Hypoxemia	0.03	0.111 ± 0.012	0.746 ± 0.093	0.118 ± 0.006	0.629 ± 0.010
Hypotension (MAP)	0.17	0.343 ± 0.006	0.539 ± 0.011	0.278 ± 0.011	0.636 ± 0.007
Hypotension (SBP)	0.06	0.200 ± 0.013	0.790 ± 0.017	0.130 ± 0.004	0.714 ± 0.006

Appendix D. Extended Experimental Results

This appendix provides extended experimental results that complement the main text. We report extended evaluations on SynSum and MIMIC-Syn, including robustness analyses and ablations, as well as additional analysis of treatment assignments on MIMIC-Real.

D.1. SynSum

Recoverability of Confounders From Text. A key assumption underlying representation-based causal methods is that the learned representations are sufficient for confounding adjustment. To assess this in SynSum, we evaluate how well the symptom variables used in the data-generating process can be recovered from the text representations. Specifically, we train classifiers to predict each symptom variable from the clinical notes. As shown in Table 10, the confounders are predictable from text, although imperfectly, confirming that the notes contain substantial confounding signal.

Table 10: **SynSum (confounder predictability from text).** Prediction of symptom confounders from clinical notes (mean ± std. over five random seeds). "Prev." denotes the positive-class prevalence in the test set. Each classifier's decision threshold was selected to maximize F1 score on the validation set.

Confounder	Prev.	F1	Acc.	AUPRC	AUROC
Cough	0.36	0.971 ± 0.001	0.979 ± 0.001	0.989 ± 0.000	0.991 ± 0.000
Dyspnea	0.20	0.949 ± 0.010	0.980 ± 0.004	0.980 ± 0.001	0.990 ± 0.001
Fever (high)	0.07	0.995 ± 0.007	0.999 ± 0.001	1.000 ± 0.000	1.000 ± 0.000
Fever (low)	0.17	0.841 ± 0.006	0.951 ± 0.002	0.879 ± 0.001	0.935 ± 0.000
Fever (none)	0.77	0.968 ± 0.001	0.950 ± 0.002	0.983 ± 0.000	0.956 ± 0.001
Nasal symptoms	0.25	0.974 ± 0.001	0.987 ± 0.000	0.984 ± 0.000	0.990 ± 0.000
Pain	0.15	0.796 ± 0.014	0.950 ± 0.003	0.851 ± 0.006	0.930 ± 0.004

Policy Performance Across Training Sizes. In addition to the main results, we report policy evaluation metrics across a wider range of training sizes (Table 11). As the number of training samples increases, all methods improve, but AACE consistently achieves better policy performance than both risk-based and representation-based causal baselines.

Table 11: **SynSum (main results).** Policy metrics across training sizes, evaluated on a fixed test set (mean ± std. dev. over five random seeds). Lower values are better. PEHE measures treatment effect estimation error and is not defined for the risk-based model. The best model within each training size is shown in **bold**. The *oracle* row reports the metrics of a policy that ranks by the true treatment effect.

Training size	Model	AUTOC (↓)	$\overline{V}$ (↓)	$V_{10\%}$ (↓)	PEHE (↓)
1,000	Risk	-0.25 ± 0.02	0.77 ± 0.01	1.17 ± 0.01	—
	DR-Emb	-0.42 ± 0.04	0.69 ± 0.01	1.14 ± 0.01	0.71 ± 0.02
	DragonNet	-0.09 ± 0.16	0.79 ± 0.04	1.21 ± 0.04	0.96 ± 0.02
	TARNet	-0.43 ± 0.03	0.69 ± 0.01	1.14 ± 0.01	0.70 ± 0.03
	AACE (ours)	$\mathbf{-0.49 \pm 0.02}$	$\mathbf{0.68 \pm 0.01}$	$\mathbf{1.12 \pm 0.01}$	$\mathbf{0.62 \pm 0.02}$
2,000	Risk	-0.23 ± 0.03	0.77 ± 0.01	1.18 ± 0.01	—
	DR-Emb	-0.46 ± 0.02	0.68 ± 0.00	1.13 ± 0.01	0.64 ± 0.02
	DragonNet	-0.10 ± 0.06	0.77 ± 0.02	1.22 ± 0.01	0.83 ± 0.04
	TARNet	-0.45 ± 0.03	0.68 ± 0.01	1.14 ± 0.01	0.67 ± 0.03
	AACE (ours)	$\mathbf{-0.53 \pm 0.01}$	$\mathbf{0.66 \pm 0.00}$	$\mathbf{1.12 \pm 0.00}$	$\mathbf{0.57 \pm 0.01}$
4,000	Risk	-0.22 ± 0.01	0.77 ± 0.00	1.18 ± 0.00	—
	DR-Emb	-0.51 ± 0.01	0.66 ± 0.00	1.12 ± 0.00	0.59 ± 0.02
	DragonNet	-0.40 ± 0.02	0.70 ± 0.00	1.14 ± 0.01	0.71 ± 0.02
	TARNet	-0.51 ± 0.02	0.66 ± 0.00	1.13 ± 0.01	0.58 ± 0.02
	AACE (ours)	$\mathbf{-0.56 \pm 0.01}$	$\mathbf{0.65 \pm 0.00}$	$\mathbf{1.11 \pm 0.01}$	$\mathbf{0.53 \pm 0.02}$
6,000	Risk	-0.21 ± 0.00	0.77 ± 0.00	1.18 ± 0.01	—
	DR-Emb	-0.52 ± 0.01	0.66 ± 0.00	1.12 ± 0.00	0.57 ± 0.01
	DragonNet	-0.45 ± 0.01	0.68 ± 0.00	1.14 ± 0.00	0.65 ± 0.01
	TARNet	-0.52 ± 0.02	0.66 ± 0.01	1.13 ± 0.00	0.57 ± 0.03
	AACE (ours)	$\mathbf{-0.57 \pm 0.01}$	$\mathbf{0.64 \pm 0.00}$	$\mathbf{1.11 \pm 0.00}$	$\mathbf{0.52 \pm 0.01}$
8,000	Risk	-0.23 ± 0.00	0.77 ± 0.00	1.18 ± 0.01	—
	DR-Emb	-0.53 ± 0.01	0.65 ± 0.00	1.12 ± 0.00	0.56 ± 0.01
	DragonNet	-0.48 ± 0.01	0.67 ± 0.00	1.13 ± 0.00	0.63 ± 0.02
	TARNet	-0.53 ± 0.00	0.65 ± 0.00	1.12 ± 0.00	0.58 ± 0.02
	AACE (ours)	$\mathbf{-0.58 \pm 0.00}$	$\mathbf{0.64 \pm 0.00}$	$\mathbf{1.11 \pm 0.00}$	$\mathbf{0.50 \pm 0.01}$
Oracle		−0.74	0.61	1.08	0.00

Extended Robustness Analysis. We complement the robustness analyses in the main text by reporting additional metrics beyond AUTOC. We first report extended results for noisy annotations, missing annotated confounders, and added irrelevant annotation variables (Tables 12–14). We then report extended results for representation masking and the alternative encoder experiment (Tables 15 and 16). Overall, these results support the main finding that AACE remains effective under moderate degradation of annotation and representation quality.

Table 12: **SynSum (noisy annotations).** Policy metrics of AACE as a function of annotation corruption, using 4,000 training samples and a fixed test set (mean ± std. dev. over five random seeds). The first column indicates the percentage of confounder annotations corrupted before training. Lower values are better.

Corruption (%)	**AUTOC** (↓)	$\overline{V}$ (↓)	$V_{10\%}$ (↓)	**PEHE** (↓)
0	-0.56 ± 0.01	0.65 ± 0.00	1.11 ± 0.01	0.53 ± 0.02
5	-0.55 ± 0.01	0.65 ± 0.00	1.11 ± 0.00	0.55 ± 0.01
10	-0.53 ± 0.01	0.66 ± 0.00	1.11 ± 0.00	0.58 ± 0.02
25	-0.49 ± 0.01	0.67 ± 0.01	1.12 ± 0.00	0.64 ± 0.02
50	-0.42 ± 0.04	0.69 ± 0.02	1.15 ± 0.01	0.78 ± 0.05
75	-0.39 ± 0.03	0.70 ± 0.01	1.15 ± 0.01	0.85 ± 0.06
100	-0.37 ± 0.06	0.70 ± 0.01	1.16 ± 0.02	0.89 ± 0.06

Table 13: **SynSum (missing relevant confounders).** Policy metrics of AACE as a function of the number of removed annotated confounders. See Table 12 for details. Lower values are better.

Dropped (n)	**AUTOC** (↓)	$\overline{V}$ (↓)	$V_{10\%}$ (↓)	**PEHE** (↓)
0	-0.56 ± 0.01	0.65 ± 0.00	1.11 ± 0.01	0.53 ± 0.02
1	-0.51 ± 0.03	0.67 ± 0.02	1.12 ± 0.00	0.60 ± 0.05
2	-0.50 ± 0.01	0.67 ± 0.01	1.12 ± 0.01	0.61 ± 0.02
3	-0.47 ± 0.04	0.68 ± 0.01	1.13 ± 0.01	0.68 ± 0.07
4	-0.41 ± 0.03	0.69 ± 0.01	1.14 ± 0.01	0.77 ± 0.03
5	-0.38 ± 0.05	0.70 ± 0.01	1.16 ± 0.01	0.88 ± 0.05

Table 14: **SynSum (irrelevant annotations).** Policy metrics of AACE as a function of the number of added irrelevant annotation variables. See Table 12 for details. Lower values are better.

Added (n)	**AUTOC** (↓)	$\overline{V}$ (↓)	$V_{10\%}$ (↓)	**PEHE** (↓)
0	-0.56 ± 0.01	0.65 ± 0.00	1.11 ± 0.01	0.53 ± 0.02
1	-0.56 ± 0.01	0.65 ± 0.00	1.11 ± 0.00	0.53 ± 0.01
2	-0.56 ± 0.01	0.65 ± 0.00	1.11 ± 0.00	0.53 ± 0.01
3	-0.56 ± 0.01	0.65 ± 0.00	1.11 ± 0.00	0.54 ± 0.01
4	-0.56 ± 0.01	0.65 ± 0.00	1.11 ± 0.01	0.53 ± 0.01
5	-0.55 ± 0.01	0.65 ± 0.00	1.11 ± 0.01	0.54 ± 0.01

Table 15: **SynSum (representation masking).** Policy metrics under random masking of embedding dimensions, using 4,000 training samples and a fixed test set (mean ± std. dev. over five random seeds). The first column indicates the percentage of embedding dimensions randomly masked for all methods. Lower values are better. PEHE measures treatment effect estimation error and is not defined for the risk-based model. The best model within each masking level is shown in **bold**. The *oracle* row reports the metrics of a policy that ranks by the true treatment effect.

Masking (%)	**Model**	**AUTOC** (↓)	$\overline{V}$ (↓)	$V_{10\%}$ (↓)	**PEHE** (↓)
0	Risk	-0.22 ± 0.01	0.77 ± 0.00	1.18 ± 0.00	—
	DR-Emb	-0.51 ± 0.01	0.66 ± 0.00	1.12 ± 0.00	0.59 ± 0.02
	DragonNet	-0.40 ± 0.02	0.70 ± 0.00	1.14 ± 0.01	0.71 ± 0.02
	TARNet	-0.51 ± 0.02	0.66 ± 0.00	1.13 ± 0.01	0.58 ± 0.02
	AACE (ours)	$\mathbf{-0.56 \pm 0.01}$	$\mathbf{0.65 \pm 0.00}$	$\mathbf{1.11 \pm 0.01}$	$\mathbf{0.53 \pm 0.02}$
75	Risk	-0.21 ± 0.01	0.77 ± 0.00	1.19 ± 0.00	—
	DR-Emb	-0.47 ± 0.01	0.68 ± 0.00	1.13 ± 0.01	0.64 ± 0.02
	DragonNet	-0.28 ± 0.05	0.74 ± 0.01	1.17 ± 0.02	0.80 ± 0.03
	TARNet	-0.46 ± 0.04	0.67 ± 0.01	1.14 ± 0.01	0.65 ± 0.06
	AACE (ours)	$\mathbf{-0.54 \pm 0.01}$	$\mathbf{0.66 \pm 0.00}$	$\mathbf{1.12 \pm 0.00}$	$\mathbf{0.56 \pm 0.01}$
90	Risk	-0.19 ± 0.01	0.78 ± 0.00	1.20 ± 0.00	—
	DR-Emb	-0.43 ± 0.01	0.69 ± 0.00	1.13 ± 0.01	0.68 ± 0.02
	DragonNet	-0.27 ± 0.04	0.75 ± 0.01	1.17 ± 0.02	0.86 ± 0.02
	TARNet	-0.44 ± 0.03	0.68 ± 0.01	1.14 ± 0.01	0.67 ± 0.06
	AACE (ours)	$\mathbf{-0.51 \pm 0.01}$	$\mathbf{0.67 \pm 0.01}$	$\mathbf{1.12 \pm 0.00}$	$\mathbf{0.60 \pm 0.01}$
95	Risk	-0.17 ± 0.02	0.78 ± 0.00	1.20 ± 0.00	—
	DR-Emb	-0.36 ± 0.03	0.71 ± 0.00	1.16 ± 0.01	0.74 ± 0.01
	DragonNet	-0.29 ± 0.02	0.75 ± 0.01	1.15 ± 0.00	0.89 ± 0.03
	TARNet	-0.39 ± 0.03	0.70 ± 0.01	1.15 ± 0.01	0.73 ± 0.02
	AACE (ours)	$\mathbf{-0.46 \pm 0.00}$	$\mathbf{0.69 \pm 0.00}$	$\mathbf{1.13 \pm 0.01}$	$\mathbf{0.65 \pm 0.01}$
100	Risk	-0.18 ± 0.01	0.78 ± 0.00	1.20 ± 0.00	—
	DR-Emb	-0.28 ± 0.02	0.76 ± 0.01	$\mathbf{1.15 \pm 0.00}$	0.93 ± 0.01
	DragonNet	-0.27 ± 0.03	0.76 ± 0.01	1.16 ± 0.01	0.95 ± 0.02
	TARNet	-0.27 ± 0.01	0.76 ± 0.00	1.15 ± 0.01	0.94 ± 0.03
	AACE (ours)	$\mathbf{-0.32 \pm 0.01}$	$\mathbf{0.74 \pm 0.00}$	1.15 ± 0.01	$\mathbf{0.75 \pm 0.00}$
Oracle		-0.74	0.61	1.08	0.00

Table 16: **SynSum (alternative encoder).** Policy metrics of all methods using Gemma embeddings, using 4,000 training samples and a fixed test set (mean ± std. dev. over five random seeds). Lower values are better. PEHE measures treatment effect estimation error and is not defined for the risk-based model. The best model is shown in **bold**.

Model	AUTOC (↓)	$\overline{V}$ (↓)	$V_{10\%}$ (↓)	PEHE (↓)
Risk	-0.25 ± 0.01	0.77 ± 0.00	1.17 ± 0.00	—
DR-Emb	-0.57 ± 0.01	0.65 ± 0.00	1.11 ± 0.00	0.53 ± 0.02
DragonNet	-0.48 ± 0.00	0.68 ± 0.00	1.13 ± 0.00	0.63 ± 0.01
TARNet	-0.54 ± 0.03	0.66 ± 0.01	1.12 ± 0.00	0.56 ± 0.04
AACE (ours)	$\mathbf{-0.60 \pm 0.00}$	$\mathbf{0.64 \pm 0.00}$	$\mathbf{1.10 \pm 0.00}$	$\mathbf{0.48 \pm 0.02}$

D.2. MIMIC-Syn

Recoverability of Confounders From Text. As in SynSum, we assess how well the confounders used in the data-generating process can be recovered from the text representations. Specifically, we train classifiers to predict the cardiovascular diagnosis variables from the clinical notes. As shown in Table 17, these confounders are predictable from text, although imperfectly, indicating that the notes contain substantial confounding signal.

Table 17: **MIMIC-Syn (confounder predictability from text).** Prediction of diagnostic confounders from clinical notes (mean ± std. over five random seeds). "Prev." denotes the positive-class prevalence in the test set. Each classifier's decision threshold was selected to maximize F1 score on the validation set.

Confounder	Prev.	F1	Acc.	AUPRC	AUROC
Atrial fibrillation	0.24	0.694 ± 0.007	0.856 ± 0.005	0.759 ± 0.004	0.889 ± 0.002
Congestive heart failure	0.22	0.591 ± 0.005	0.794 ± 0.009	0.584 ± 0.004	0.845 ± 0.002
Coronary atherosclerosis	0.26	0.766 ± 0.003	0.884 ± 0.006	0.854 ± 0.004	0.917 ± 0.002
Hypertension	0.44	0.648 ± 0.002	0.573 ± 0.009	0.615 ± 0.002	0.694 ± 0.002

D.3. MIMIC-Real

Additional Analysis of Treatment Assignments. Since ground-truth treatment effects are unavailable in MIMIC-Real, we cannot determine which method produces the best policy. We therefore restrict the analysis to agreement between the treatment rankings induced by the different methods, and between AACE and the observed treatments. Table 18 reports the corresponding Jaccard overlaps and Spearman rank correlations. The results show that AACE is more strongly aligned with the representation-based causal baseline than with the risk-based model, especially in terms of global rank correlation. At the same time, agreement between AACE and the observed treatment assignments is low, indicating that the learned rankings differ substantially from the treatment decisions

recorded in practice. Overall, these results reinforce the main paper's qualitative conclusion that annotation-assisted, risk-based, and representation-based causal strategies can prioritize different patient groups on real-world EHR data.

Table 18: **MIMIC-Real (policy overlap and agreement).** Policy overlap and ranking agreement on real-world data. Jaccard measures agreement between top-k treatment assignments, while Spearman ρ captures global rank correlation. The final column compares the learned ranking of AACE with observed treatment assignment and is reported separately because observed treatment is binary rather than a full ranking.

Metric	**AACE vs TARNet**	**AACE vs Risk**	**Risk vs TARNet**	**AACE vs Observed**
Jaccard overlap				
$k = 5\%$	0.088 ± 0.106	0.037 ± 0.027	0.316 ± 0.166	0.034 ± 0.004
$k = 10\%$	0.155 ± 0.115	0.076 ± 0.048	0.321 ± 0.170	0.055 ± 0.006
$k = 20\%$	0.245 ± 0.088	0.141 ± 0.052	0.340 ± 0.148	0.078 ± 0.022
$k = 50\%$	0.573 ± 0.075	0.362 ± 0.051	0.421 ± 0.126	0.146 ± 0.014
Spearman correlation				
ρ	0.551 ± 0.202	0.067 ± 0.170	0.233 ± 0.313	0.070 ± 0.057

Appendix E. Experimental Details

This section summarizes implementation details common to all experiments. We use the same architecture, tuning procedure, and early stopping strategy across datasets and model components. All models (classifiers, nuisance models, and treatment effect regressors) were implemented as small neural networks with a single hidden layer and ReLU activations, using a sigmoid output layer for binary classification tasks and a linear output layer for regression tasks. Models were trained with Adam, optional weight decay, and a maximum of 50 epochs, with early stopping based on validation loss (patience of five epochs). The confounder recoverability experiments used class-weighted binary cross-entropy, with positive class weight $(1-p)/p$, where p is the positive-class prevalence in the training set, other classification tasks used standard binary cross-entropy, and regression tasks used mean squared error. For the doubly robust pseudo outcomes, we do not apply cross-fitting. Hyperparameters were tuned once per model family on a single training size and then reused across experiments. We performed a random search with 12 trials over the following grid: hidden dimension $\in \{64, 128\}$, learning rate $\in \{10^{-4}, 3 \times 10^{-4}, 5 \times 10^{-4}, 10^{-3}\}$, weight decay $\in \{0, 10^{-5}, 10^{-4}\}$, and batch size $\in \{128, 256\}$. The configuration with the lowest validation loss was selected. All experiments used 10% of the data for validation and 10% for testing, with the remaining data used for training. Random seeds affected both data splits and weight initialization.