Preference-based Conditional Treatment Effects and Policy Learning

Dovid Parnas*
Technion,
Israel Institute of Technology

Mathieu Even*
Inria, Inserm,
Université de Montpellier,
France

Julie Josse
Inria, Inserm,
Université de Montpellier
France

Uri Shalit
Department of Statistics
and Operations Research,
Tel Aviv University

Abstract

We introduce a new preference-based framework for conditional treatment effect estimation and policy learning, built on the Conditional Preference-based Treatment Effect (CPTE). CPTE requires only that outcomes be ranked under a preference rule, unlocking flexible modeling of heterogeneous effects with multivariate, ordinal, or preference-driven outcomes. This unifies applications such as conditional probability of necessity and sufficiency, conditional Win Ratio, and Generalized Pairwise Comparisons. Despite the intrinsic non-identifiability of comparison-based estimands, CPTE provides interpretable targets and delivers new identifiability conditions for previous unidentifiable estimands. We present estimation strategies via matching, quantile, and distributional regression, and further design efficient influence-function estimators to correct plug-in bias and maximize policy value. Synthetic and semi-synthetic experiments demonstrate clear performance gains and practical impact.

1 INTRODUCTION

In high-stakes domains, such as healthcare and public policy, clinicians and policymakers must make decisions about treatments whose effects can vary widely between individuals. Population-level averages are rarely enough: the right choice for each person depends on their specific characteristics, risk factors and likely response. To truly personalize data-driven decisions, we must estimate individual effects and learn feature-dependent policies.

The standard approach estimates conditional treatment effects, contrasting conditional outcomes for each treatment arm, and derives policies accordingly. However, while the conditional average treatment effect (CATE) has become the default tool for heterogeneous effect estimation and policy learning, it can be misleading. CATE-based policies rely on averages, ignoring complex treatment responses. For example, a heavy tailed distribution can make CATE-based policies favor treatments where the majority actually loses. Consider salary as an outcome and an intervention that causes small monthly losses for many low-salary individuals but large gains for a handful of high-salary individuals. A perfectly estimated CATE would recommend the intervention, despite diminishing the welfare of the majority. In critical settings, this bias is not just a statistical artifact, it threatens fair, effective decision-making.

Moreover, defining a CATE policy is not straightforward for multiple outcomes. Consider heart-failure studies (Pocock et al., 2011) where death, stroke, and hospitalization are all outcomes of primary interest. These outcomes follow a natural hierarchy: death being the most severe, followed by stroke and then hospitalization. The challenge lies in incorporating this hierarchy into a principled, data-driven framework for decision making. A leading approach over the past decade is the *Win Ratio* (Pocock et al., 2011), where treated and control patients are paired and compared sequentially. Death is assessed first; stroke is compared if both survived; if neither patient experienced a stroke, hospitalization is evaluated. This design explicitly respects clinical priorities. However, quantifying treatment effects, conditional effects, and learning policies when outcomes can only be judged through comparisons remains highly non-trivial (Mueller and Pearl, 2023; Even and Josse, 2025; Tian and Pearl, 2000). They cannot be captured by CATE. The core

Proceedings of the 29th International Conference on Artificial Intelligence and Statistics (AISTATS) 2026, Tangier, Morocco. PMLR: Volume 300. Copyright 2026 by the author(s).

difficulty is the non-identifiability of the underlying estimands; it is impossible to estimate them from observable data. due to the *fundamental problem of causal inference* (Holland, 1986). Indeed, access to the *joint* potential outcomes distribution is required to compare counterfactuals. However, no unit can ever be observed in both the treatment and control groups, even in randomized studies.

This paper introduces a general framework for learning preference-based conditional treatment effects and policies from both randomized and observational data. Preference-based learning directly answers the question, "Which treatment should be preferred?" It also expands policy treatment effect estimation and policy learning to outcomes that are comparable but not quantifiable. A first example is the probability of necessity and sufficiency (PNS), a classic causal quantity (Mueller and Pearl, 2023; Tian and Pearl, 2000). PNS can be interpreted as the probability that the treated outcome is strictly preferred to the control outcome for a given individual, averaged across the population. Although related directions have been explored (Wang and Li, 2025; Wang et al., 2025; Shingaki et al., 2021), it lacks an identifiable, conditional construct tailored to personalized decision-making.

More broadly, preference-based effects extend naturally to complex, multivariate outcomes whenever a rule specifies which outcome is preferred in a pairwise comparison. This framework subsumes hierarchical outcome settings such as the Win Ratio and Generalized Pairwise Comparisons (Pocock et al., 2011; Buyse, 2010). Our method goes beyond these cases, learning conditional effects and optimal policies under such preference structures, while also accommodating more general, possibly non-strict, preference rules.

Related work. Most work on conditional treatment effects and policy learning in causal inference focuses on single real-valued or binary outcomes (Athey and Imbens, 2016; Wager et al., 2019; Künzel et al., 2019; Athey and Wager, 2021). This paper offers a broader framework, connecting three related lines of work: *(i)* non-identifiable causal quantities such as the PNS and their bounds, *(ii)* pairwise comparison methods like the Win Ratio and Net Benefit, and *(iii)* distributional regression and quantile treatment effects.

The *Probability of Necessity and Sufficiency* (*PNS*) is the probability that treatment improves the outcome and is required to do so (Mueller and Pearl, 2023; Tian and Pearl, 2000). It can be viewed as the population average of individual-level effects, with subpopulation extensions studied in Kawakami et al. (2024); Wang and Li (2025). Lei (2024) presents *population*-level point estimates of PNS. However, to our knowledge, a clear definition of an identifiable, conditional PNS is missing.

Because PNS and its conditional variants are fundamentally *non-identifiable*, most prior work has focused on deriving identifiable *bounds* from data under suitable assumptions (Shingaki et al., 2021; Robins and Richardson, 2010; Tian and Pearl, 2000). Bounding strategies have been developed for binary (Mueller and Pearl, 2023), ordinal (Huang et al., 2017; Lu et al., 2018; Li and Pearl, 2024; de Aguas et al., 2025), continuous (Manski, 2003), quantile treatment effects (QTE) (Koenker and Bassett, 1978b; Firpo, 2007) and conditional QTE (Abadie et al., 2002; Firpo, 2007; Cui and Han, 2023). Instead of relying on identifiable bounds, which can be loose for conditional quantities or non-binary outcomes (Li and Pearl, 2024; Manski, 2003), we introduce and estimate an interpretable, identifiable proxy.

Pocock et al. (2011) and Buyse (2010) introduced the *Win Ratio* and *Net Benefit* to compare two treatments over multiple, *hierarchical* outcomes. These metrics allow several outcomes to factor into the evaluation of treatment effects and decision making. Petit et al. (2025) considered optimal treatment rules for the net benefit. However, our broader framework goes further by defining conditional treatment effect with hierarchical outcomes, enabling policy value optimization and developing semi-parametric efficient estimators. Our work borrows the formalism of (Even and Josse, 2025) and non-causal Probabilistic Index Models (PIMs, (Thas et al., 2012)).

Our estimation methods are based on nearest neighbor approaches (Biau and Devroye, 2015) and off-the-shelf quantile regression methods (Koenker and Bassett, 1978a; Greene, 2003; Meinshausen and Ridgeway, 2006) for nuisance parameters. Our methods differ from distributional treatment effect approaches (Näf and Susmann, 2024; Zhou et al., 2022; Park et al., 2021; Kallus and Oprescu, 2023): we use distribution methods *as a means* to estimate preference-based effects.

Outline of the paper and contributions. We first introduce the causal and preference-based setting in Section 2 and define our target estimands in Section 3. We distinguish between a non-identifiable conditional Individual Treatment Effect (ITE) and our new identifiable estimand, the Conditional Preference Treatment Effect (CPTE), and establish conditions under which they coincide via the concept of *structural treatment effect modifiers*. In Section 4, we develop estimation strategies for CPTE, using flexible building blocks like matching and distributional regression. In Section 5, we turn to preference-based policy learning, introduc-

ing a plug-in method and a policy value optimization approach. Finally, in Section 6, we derive a semiparametric efficient policy estimator based on a one-step efficient influence function (EIF) correction. We validate these methods on synthetic and semi-synthetic experiments (Section 7, Appendix E). Our contribution is threefold, enabling practical policy learning within a preference-based framework. We build an identifiable conditional proxy, the CPTE, for individual-level preference and establish its identifiability conditions. We develop an estimation procedure for the CPTE. Finally, we learn decision policies based on preference rules.

2 SETTING

Causal assumptions. We assume access to n i.i.d. samples, associated with feature vectors $X_i \in \mathcal{X}$, treatment $T_i \in \{0,1\}$, and observed outcome $Y_i \in \mathcal{Y}$ (possibly multivariate, e.g. $\mathcal{Y} \subset \mathbb{R}^d$). Under the *potential outcome* framework (Splawa-Neyman et al., 1990), each unit has two outcomes, $Y_i(0), Y_i(1) \in \mathcal{Y}$, only one of which is observed.

We assume *SUTVA*, *unconfoundedness* and *positivity*:

Assumption 1 (SUTVA)**.** *$Y_i = Y_i(T_i)$ for all $i \in [n]$.*

Assumption 2 (unconfoundedness)**.** *$\{Y_i(0), Y_i(1)\} \perp T_i \mid X_i$.*

Assumption 3. *There exists $\eta \in (0,1)$ such that $\eta \leqslant e(x) \leqslant 1-\eta$ for all $x \in \mathcal{X}$, where $e(x) = \mathbb{P}(T_i = 1 \mid X_i = x)$.*

Preference-based framework. We ask whether a unit would fare better under treatment than without. Given two outcomes $y, y' \in \mathcal{Y}$, "better" cannot always be quantified correctly with the sign and magnitude of their linear difference $y - y'$. Instead, we consider a more general "preference function":

Definition 1 (Preference function)**.** *$w : (y, y') \in \mathcal{Y}^2 \mapsto w(y|y') \in \mathbb{R}$ quantifies how y fares against y'.*

Typical examples of preference functions are:
(i) PNS, For single real-valued outcomes:

$$w(y|y') = \mathbf{1}\{y > y'\}\,.$$

(ii) General Ordered Outcomes like the Win Ratio:

$$w(y|y') = \mathbb{1}_{\{y \succ y'\}} + 0.5 \cdot \mathbb{1}_{\{y \sim y'\}}\,, \tag{1}$$

where $\succ$ is an order on $\mathcal{Y}$, such as the *lexicographic order* on $\mathbb{R}^d$: $y \succ y'$ iff $\inf\{k : y_k > y'_k\} < \inf\{k : y_k < y'_k\}$ used for hierarchical outcomes (Pocock et al., 2011).

(iii) Unknown preferences: If an "ideal stochastic oracle" compares the outcomes Y and Y' of two units,

$$w(Y|Y') \stackrel{\text{def}}{=} \mathbb{P}(Y \succ Y') + \tfrac{1}{2}\mathbb{P}(Y \sim Y'),$$

where probabilities capture uncertainty or individual preferences. In this paper, w is assumed to be known.

3 CONDITIONAL EFFECTS

3.1 Definition of a preference-based conditional treatment effect

Personalized treatment is the holy grail of causal inference. While population-level effects inform broad policy decisions, they overlook the substantial heterogeneity that individualized policies can account for. However, the Individual Treatment Effect (ITE) is inaccessible due to the *fundamental problem of causal inference*. In our preference-based framework with preference function w, the ITE for patient i is:

$$\Delta_i = w(Y_i(1)|Y_i(0))\,.$$

It measures how $Y_i(1)$ compares to $Y_i(0)$; treatment follows Δ_i. Because only one action is observed, the ITE is in general unidentifiable and cannot be estimated from data. Common surrogates for ITE are *conditional treatment effects*, which in many cases can be estimated from data. Conditional effects answer subgroup causal questions and support policy learning (Athey and Wager, 2021), ideally balancing identifiability and the ability to tailor treatments to specific populations. For the risk difference (RD) $w(y|y') = y - y'$, the widely used Conditional Average Treatment Effect (CATE) is exactly the conditional mean of the ITE:

$$\tau_{\text{RD}}(x) = \mathbb{E}\left[Y_i(1) - Y_i(0) | X_i = x\right]\,.$$

It is identifiable under Assumptions 1 to 3 and estimable via a variety of methods (Nie and Wager, 2021; Wager and Athey, 2018; Künzel et al., 2019). Unfortunately, for the more generalized preference functions considered in our paper, we have the following unidentifiability result.

Lemma 1. *The conditional ITE, defined as:*

$$\mathbb{E}\left[\Delta_i | X_i = x\right] = \mathbb{E}\left[w(Y_i(1)|Y_i(0)) \mid X_i = x\right]\,, \tag{2}$$

is generally unidentifiable and cannot be inferred from the distribution of the observations $(X_i, A_i, Y_i)_{i \in [n]}$.

This stems from the non-separability of w, e.g. $w(y|y') \neq g(y) - g(y')$, see proof in Appendix B.1. When w is separable and can be written as $w(y|y') = g(y) - g(y')$ for some function g, this quantity is the classical CATE with $g(Y_i)$ as outcome, and is identifiable. Note that the non-identifiability in Lemma 1 does *not* stem from confounding and holds even in RCT settings. Mueller and Pearl (2023); Cui and Han

(2023); Li and Pearl (2024) explore identifiable bounds on the conditional ITE for the PNS and QTE.

We propose overcoming the non-identifiability of the non-separable, conditional ITE by defining an *identifiable* conditional effect. We call our new estimand the Conditional Preference Treatment Effect (CPTE). It is obtained by conditioning on covariates and replacing the joint distribution of counterfactuals, $(Y_i(1), Y_i(0))|X_i = x$, with $(Y_i(1), Y_j(0))|(X_i, X_j) = (x, x)$, drawing inspiration from the formalism of probabilistic index models (Thas et al., 2012). In the next subsection, we introduce identifiability conditions under which the CPTE will be equal to the Conditional ITE.

Definition 2 (Preference-based conditional effect)**.** *For $x \in \mathcal{X}$, let $\{Y(0,x), Y(1,x)\}$ be an independent copy of $\{Y_i(0), Y_i(1)\}|X_i = x$. Define*

$$\begin{aligned}\mathrm{CPTE}(x) &\overset{def}{=} \mathbb{E}\left[w(Y(1,x)|Y_i(0)) \,|\, X_i = x\right] \\ &= \mathbb{E}\left[w(Y_i(1)|Y(0,x) \,|\, X_i = x\right] ,\end{aligned} \tag{3}$$

the Conditional Preference Treatment Effect. Equivalently, for $i \neq j$,

$$\mathrm{CPTE}(x) = \mathbb{E}\left[w(Y_i(1)|Y_j(0))|(X_i, X_j) = (x, x)\right] ,$$

where the expectation is over the products of the probability distributions $Y_i(1)|X_i = x$ and $Y_j(0)|X_j = x$.

Under Assumptions 1 to 3,

$$\mathrm{CPTE}(x) = \mathbb{E}\left[w(Y_i|Y_j) \,|\, X_i = X_j = x, T_i = 1, T_j = 0\right]$$

Therefore, it is *identifiable*. A population effect can be defined as $\mathbb{E}\left[\mathrm{CPTE}(X_i)\right]$, which is *directly collapsible* (written as the population average of conditional effects (Even and Josse, 2025; Colnet et al., 2023)). Note that because an independent copy always exists, we do not require further assumptions for the definition of $\mathrm{CPTE}(x)$, unlike the net benefit as defined by Petit et al. (2025). Finally, we define treatment effect modifiers as follows:

Definition 3. *A variable X^k is a treatment effect modifier if $u \mapsto \mathbb{E}\left[\mathrm{CPTE}(X_i)|X_i^k = u, X_i^{-k} = v\right]$ is not constant when conditioning on other features X^{-k}. Effects are homogeneous when no modifier exists, and heterogeneous otherwise.*

Treatment-effect modifiers depend on the effect measure w and may differ from those of the classical risk difference. Next, we explain how our CPTE relates to the non-identifiable Conditional ITE $\mathbb{E}\left[\Delta_i|X_i = x\right]$. In particular, the next subsection provides new identifiability conditions under which $\mathrm{CPTE}(x) = \mathbb{E}\left[\Delta_i|X_i = x\right]$, as well as tools to assess how far $\mathrm{CPTE}(x)$ is from the non-identifiable Conditional ITE when these conditions do not hold.

3.2 Properties of $\mathrm{CPTE}(x)$ and dependencies over features

Which covariates should be included in X_i? In order for $\mathrm{CPTE}(x)$ to be identifiable, the classic causal assumptions (Assumptions 1 to 3) must hold. In particular, Assumption 2 requires that all confounders must be included as covariates. *Treatment effect modifiers* (Definition 3) are features that have an influence on the treatment effect. However, they are not necessarily confounders, that is, variables that also influence treatment assignment. Nevertheless, $\mathrm{CPTE}(x)$ requires them to capture population heterogeneity, as is often the case for conditional effects.

To achieve identifiability of the Conditional ITE, defined in Equation (2), *via* the equality $\mathrm{CPTE}(x) = \mathbb{E}\left[\Delta_i|X_i = x\right]$, treatment effect modifiers and confounders are not enough, as highlighted in Lemma 1. Bearing this in mind, we introduce the notion of *structural treatment effect modifiers*. These are all the variables that could make the outcome of a treated patient better or worse than that of a control, from the set of variables Z_i that spans all endogenous and exogenous variables in the structural causal model, as introduced by Pearl (2009, Chapter 1). This definition differs from that of Definition 3, which only considers *observed* variables.

Definition 4 (Structural treatment effect modifiers)**.** *Structural treatment effect modifiers are a minimal set of variables $\mathcal{I} \subset [p]$ for which there exists a non-constant function g such that for all $i, j \in [n]$ and all $x \in \mathbb{R}^{\mathcal{I}}$:*

$$\mathbb{E}\left[w(Y_i(1), Y_j(0)) \,|\, Z_i, Z_j, (Z_i^{\mathcal{I}}, Z_j^{\mathcal{I}}) = (x, x)\right] = g(x) ,$$

where $z^{\mathcal{I}} = (z_k)_{k \in \mathcal{I}}$ for $z \in \mathbb{R}^p$.

As the next lemma shows, if all structural treatment effect modifiers are included in the covariates, then we have equality between $\mathrm{CPTE}(x)$ and $\mathbb{E}\left[w(Y_i(1)|Y_i(0))|X_i = x\right]$. However, if this does not hold, $\mathrm{CPTE}(x)$ can function as an identifiable proxy for $\mathbb{E}\left[w(Y_i(1)|Y_i(0))|X_i = x\right]$, requiring *sensitivity analyses* to help assess how far we are from equality and identifiability (Nguyen et al., 2018; Colnet et al., 2022; Jin et al., 2023).

Theorem 1. *Assume that structural treatment effect modifiers in the sense of Definition 4 are included in the set of features X_i. Then, $\mathrm{CPTE}(x) = \mathbb{E}\left[w(Y_i(1)|Y_i(0))|X_i = x\right]$ for all $x \in \mathcal{X}$.*

In summary, features X_i need to include *confounding variables* for identifiability of the CPTE (in the case of an observational study), *treatment effect modifiers* to capture treatment effect heterogeneity, and *structural treatment effect modifiers* for the equality in The-

orem 1 to hold. However, it might not always be possible to include, or even know, all structural treatment effect modifiers. For example, in the medical context, these can come from genetic factors or other unmeasured sources of variation. Yet, as we show in the experiments below, even in such cases the CPTE can be a useful estimand. We note that another assumption which leads to the same identifiability result is *potential independence*, namely that $Y_i(1) \perp\!\!\!\perp Y_i(0)|X_i$ for all i. This assumption, stating that the potential outcome under control tells nothing about the potential outcome under treatment, is unlikely to hold.
A presentation of the above definitions and Theorem 1 in SCM terminology is available in the Appendix A.2.

3.3 Examples

Preference conditional PNS. For $\mathcal{Y} \subset \mathbb{R}$ and $w(y|y') = \mathbb{1}_{\{y>y'\}}$, we obtain a conditional and identifiable PNS that we call *preference conditional PNS*. This contrasts with the classical PNS (Tian and Pearl, 2000) and its conditional counterparts, which are unidentifiable. Conditioning on covariates and comparing treated and untreated outcomes across independent copies yields an identifiable estimand, preserving the causal meaning of the PNS and enabling personalized policies. With *binary outcomes*, since a Bernoulli distribution is determined by its mean, CPTE(x) is a direct function of $\mathbb{E}[Y_i|X_i = x, A_i = a]$, estimable via standard machine learning. However, for general real-valued outcomes or for the non-identifiable PNS, no closed form exists. In Section 4 we show how distributional regression methods can be used to estimate the CPTE in this case.

Conditional effects with hierarchical outcomes. When multiple outcomes, $\mathcal{Y}$, have a hierarchical order $\succ$ (e.g. $\mathcal{Y} \subset \mathbb{R}^d$ with a lexicographic order) and $w(y|y') = \mathbb{1}_{\{y\succ y'\}}$, treatment effects are often estimated by Win Ratio or pairwise comparisons. These usually target the population-level measure $\tau_{\text{pop}} = \mathbb{E}[w(Y_i(1)|Y_j(0))]$ (Mao, 2017) for independent i, j. Such measures quantify the preference function of a treated unit against a control unit when both are sampled at random from the population distribution. By conditioning on $(X_i, X_j) = (x, x)$, CPTE(x) becomes a personalized version, facilitating policy learning for complex outcomes.

Conditional effects with unknown preferences. When preferences are unknown, w must be learned. A common approach posits a latent reward R with $w(y|y') = \mathbb{E}[\exp(R(y) - R(y'))]$, estimated from expert or participant feedback. One may also allow R to depend on X, yielding unit-specific preferences. Then CPTE$(x) = \mathbb{E}\left[e^{R(Y_i(1))}|X_i = x\right]\mathbb{E}\left[e^{-R(Y_i(0))}|X_i = x\right]$, where w (and R) is assumed learned.

4 CPTE ESTIMATION METHODS

We now turn to estimating $x \mapsto$ CPTE(x) defined in Definition 2. CPTE captures more than first-moment information of the conditional distributions of the counterfactuals $(\mathbb{E}[Y_i(t)|X_i = x])$. Therefore, our estimation methods learn the conditional distributions $Y_i(1)|X_i$ and $Y_i(0)|X_i$. If we know how to sample from $\mathcal{P}_{Y_i|T_i,X_i}$, a consistent and asymptotically normal estimator of CPTE(x) would be:

$$\frac{1}{m}\sum_{k=1}^{m} w(\hat{y}_1^{(k)}, \hat{y}_0^{(k)}), \tag{4}$$

where $(\hat{y}_t^{(k)})_k$ are independent random variables sampled from $\mathcal{P}_{Y_i|T_i=t,X_i=x}$, with $(\hat{y}_1^{(k)}) \perp\!\!\!\perp (\hat{y}_0^{(k)})$. This will form the basis of our general approach. We offer several methods of utilizing the available data, (X_i, T_i, Y_i), to sample from $\mathcal{P}_{Y_i|T_i=t,X_i=x}$ and estimate eq. (4).

4.1 Matching Approach

We first propose a matching approach, for which we prove consistency under Assumptions 1 to 3 and an additional regularity assumption. We use a $k-$nearest neighbor approach that yields non-parametric consistent estimators for CPTE(x). Let $\mathcal{N}_t \subset [n]$ be the set of patients with treatment assignment $T_i = t$. For $x \in \mathcal{X}$ and $k \in [n]$, let the $k-$neighborhood of x in arm $t \in \{0, 1\}$ be denoted as $\mathcal{N}_t(x, k)$, and defined as the minimizer of $\sum_{i\in\mathcal{I}} \|x - X_i\|^2$ over all sets $\mathcal{I} \subset \mathcal{N}_t$ of size k. Let, for any $x \in \mathcal{X}$:

$$\hat{q}_W(x) \stackrel{\text{def}}{=} \frac{1}{k^2}\sum_{(i,j)\in\mathcal{N}_1(x,k)\times\mathcal{N}_0(x,k)} w(Y_i|Y_j),$$

be the distributional $k-$NN estimator, satisfying the following:

Theorem 2. *Assume that Assumptions 1 to 3 hold. If the function $x, x' \mapsto \mathbb{E}[w(Y_i(1)|Y_j(0))|X_i = x, X_j = x']$ is continuous, for all $x \in \mathcal{X}$, then in probability:*

$$\hat{q}_W(x) \longrightarrow q_W \stackrel{def}{=} \text{CPTE}(x),$$

as long as $k \to \infty$ and $\frac{k\log(n)}{n} \to 0$, where q_W stands for the CPTE.

The proof is in Appendix C. For $x \in \mathcal{X}$ and $(i, j) \in \mathcal{N}_1(x, k) \times \mathcal{N}_0(x, k)$, $w(Y_i|Y_j)$ can be interpreted as an approximate sample of $\mathcal{P}_{Y_i|X_i=x,T_i=1} \times \mathcal{P}_{Y_j|X_j=x,T_j=0}$, which can be used to calculate $\hat{q}_W$ with Equation (4).

We can also define $\hat{q}_L$ as $w_L(y|y') = w(y'|y)$ for the 'anti-preference'. Like the 'preference' function $q_W(x)$, $\hat{q}_L(x) \to q_L(x)$ with distributional K-NN:

$$q_L(x) \stackrel{\text{def}}{=} \mathbb{E}\left[w(Y_i(0), Y_j(1))|X_i = X_j = x\right] . \quad (5)$$

4.2 Quantile Regression Approach

Quantile regression methods (Koenker and Bassett, 1978a) can be used to estimate both parametrically or non-parametrically the inverse CDF of $Y(1, x)$ and $Y(0, x)$. This allows us to draw samples from the estimated inverse CDF which can be used to estimate $\mathbb{E}\left[w(Y_i(0), Y_j(1))|X_i = X_j = x\right]$. We present in our experiments results with linear quantile regression (Seabold and Perktold, 2010; Greene, 2003) and random forest quantile regression (Meinshausen and Ridgeway, 2006). For any given quantile model we can use the general method described in Algorithm 1 (Appendix A.4) to sample from the inverse CDF, giving us *distributional* regression methods. Algorithm 1 can also be used as a basis for multivariate outcomes if Y is $d-$dimensional by using the Rosenblatt transform (Rosenblatt, 1952).

5 PREFERENCE-BASED OPTIMAL POLICIES

5.1 Policy Value

A policy is a function $\pi : \mathcal{X} \to \{0, 1\}$ that maps features to treatment assignment. With real-valued outcomes, an optimal policy is a policy $\pi^\star$ that maximizes the policy value, $\mathbb{E}\left[Y_i(\pi(X_i))\right]$, over all policies π in a policy class $\Pi \subset \{0, 1\}^{\mathcal{X}}$.

How is a policy's preference to be measured? We define the **'Preference' Policy Value**, which quantifies how preferable the treatments assigned by the policy are to the alternative. For example a clinical trial may wish to estimate how preferable treating all units with treatment is to control. Let the **'Preference' Policy Value** write as, for any $\pi : \mathcal{X} \to \{0, 1\}$:

$$V(\pi) \stackrel{\text{def}}{=} \mathbb{E}\left[\pi(X_i)q_W(X_i) + (1 - \pi(X_i))q_L(X_i)\right] , \quad (6)$$

where $q_W(x) = \text{CPTE}(x)$ and $q_L(x)$ are as in Equation (5). The goal of policy learning is then to learn a policy π that maximizes "preference" over a given policy class $\Pi \subset \{0, 1\}^{\mathcal{X}}$ (Athey and Wager, 2021):

$$\pi^\star \in \operatorname*{argmax}_{\pi \in \Pi} V(\pi) . \quad (7)$$

We next propose different strategies for estimating $\pi^\star$.

5.2 Optimal Treatment Rule

An optimal treatment rule (OTR) is a policy that solves Equation (7). For unconstrained policy learning (*i.e.*, $\Pi = \{0, 1\}^{\mathcal{X}}$ is the set of all possible policies), $\pi^\star$ has a closed form expression:

$$\forall x \in \mathcal{X} , \quad \pi^\star(x) = \mathbb{1}_{\{\delta(x)>0\}} , \quad (8)$$

where:

$$\forall x \in \mathcal{X} , \quad \delta(x) \stackrel{\text{def}}{=} q_W(x) - q_L(x) . \quad (9)$$

The optimal policy recommends a treatment if the 'preference', $q_W(x)$ is higher than the 'anti-preference', $q_L(x)$. $\delta(x)$ can also be interpreted as a *conditional net benefit* of the treatment, usually defined over the whole population for hierarchical outcomes (Buyse, 2010).

Direct Plug-In Estimation of the OTR. Our first policy learning approach in estimating the OTR $\pi^\star$ defined in Equation (8), consists in directly estimating δ defined in Equation (9). With $\hat{q}_W$ and $\hat{q}_L$ (non-parametric) estimators of q_W, q_L, let $\hat{\delta}(x) \stackrel{\text{def}}{=} \hat{q}_W(x) - \hat{q}_L(x)$. The estimated OTR then writes as:

$$\hat{\pi}_{\text{OTR}}(x) \stackrel{\text{def}}{=} \mathbb{1}_{\left\{\hat{\delta}(x)>0\right\}} . \quad (10)$$

Depending on the precision with which the nuisance parameters q_W and q_L are estimated, we obtain consistent estimators of $\pi^\star$. In particular, if $\hat{q}_W, \hat{q}_L$ are built with $k-$NN, Theorem 2 guarantees that $\hat{\pi}$ is a consistent estimator of $\pi^\star$.

5.3 Plug-in Policy Value Optimization

Instead of directly plugging-in estimators of δ, q_W or q_L in the closed-form expression of the OTR $\pi^\star$ defined in Equation (8), one could consider a more restrictive policy class Π to optimize over. This would require estimating the value of a policy (defined in Equation (6)). A plug-in approach to estimate the value of a policy π would be to estimate nuisance parameters $\hat{q}_W, \hat{q}_L$, and to estimate $V(\pi)$ by $\hat{V}(\pi)$, defined as:

$$\hat{V}(\pi) \stackrel{\text{def}}{=} \frac{1}{n} \sum_{i=1}^{n} \pi(X_i)\hat{q}_W(X_i) + (1 - \pi(X_i))\hat{q}_L(X_i) . \quad (11)$$

The plug-in optimization approach then considers the following approximate optimization problem:

$$\hat{\pi}_{\text{plug-in}} \in \operatorname*{argmax}_{\pi \in \Pi} \hat{V}(\pi) . \quad (12)$$

The class Π to consider can be as follows.

Weighted classification. In our experiments, we parameterize a linear classifier as f_θ for $\theta \in \Theta$ and predict the sign of $\hat{q}_W(X_i) - \hat{q}_L(X_i)$ weighted by the magnitude of the difference, implicitly maximizing $\hat{V}(\pi)$ over

linear classifiers.

Policy trees. Policytree (Sverdrup et al., 2020) can also be used, and directly implements Equation (12), for Π the set of all binary decision trees, of a maximum depth D (specified by the user).

6 THE 1-STEP CORRECTED POLICY

Plug-in estimators for the nuisance functions in Equation (11) can cause large bias in $\hat{V}(\pi)$ if the models have large non-asymptotic biases (e.g. nearest-neighbors), if they overfit (e.g. random forests), or if they are misspecified (e.g. linear models). Efficient Influence Functions (EIFs) can be used to correct for this plug-in bias (Hines et al., 2022) and to obtain estimators that satisfy, under mild regularity conditions, double robustness properties and semiparametric efficiency. Here, efficiency is understood relative to a fully nonparametric statistical model, in the sense that EIF-based estimators achieve the lowest possible asymptotic variance among all regular, asymptotically linear estimators in this model. The existence of an efficient influence function further requires that the preference-based policy value parameter be pathwise differentiable with respect to the observed data distribution. As in classical policy value estimation, this holds under standard smoothness conditions on the parameter mapping and regularity of the preference function w. Positivity ensures that the resulting EIF is well-defined and square-integrable, and that EIF-based estimators are regular and attain the semiparametric efficiency bound in a fully nonparametric model. Therefore, we compute an EIF $\phi(\mathcal{P})(o)$ for any distribution $\mathcal{P}$ and observation o for preference policy learning. Details on and intuition behind EIF correction are presented in Appendix A.3.

Proposition 1. *Let $\psi_\pi(\mathcal{P}) = V(\pi)$ be our estimand. For $Y(t,x)$ as defined in Definition 2, let:*

$$p_W(x,t,y) = t\,\mathbb{E}\left[w(y \mid Y(0,x))\right] + (1-t)\,\mathbb{E}\left[w(Y(1,x) \mid y)\right] .$$

and similarly $p_L(x,t,y)$ by replacing w by $w_L(y|y') = w(y'|y)$. Let $e(x) = \mathbb{P}(T_i = 1|X_i = x)$ and q_W, q_L as previously defined. Then the EIF of ψ_π for a distribution $\mathcal{P}$ and $o = (X,T,Y)$ is:

$$\begin{aligned}\Phi_\pi(\mathcal{P})(o) &= -\psi_\pi(\mathcal{P}) + \pi(X)q_W(X) \\ &+ (1-\pi(X))q_L(X) \\ &+ \frac{T}{e(X)}\Big(\pi(X)(p_W(X,T,Y) - q_W(X)) \\ &\quad + (1-\pi(X))(p_L(X,T,Y) - q_L(X))\Big) \\ &+ \frac{1-T}{1-e(X)}\Big(\pi(X)(p_W(X,T,Y) - q_W(X)) \\ &\quad + (1-\pi(X))(p_L(X,T,Y) - q_L(X))\Big) .\end{aligned}$$

The proof is given in Appendix D. Let $\hat{e}$ be an estimator of the propensity score, $\hat{q}_W, \hat{q}_L$ and $\hat{p}_W, \hat{p}_L$ estimators of the nuisance parameters q_W, q_L, p_W, p_L. The 1-step policy value estimator is the first-order correction of $\hat{V}(\pi)$ (Equation (6)), and writes as $\Psi^\pi_{\text{1-step}} \overset{\text{def}}{=} \Psi^\pi_{\text{plug-in}} + \frac{1}{n}\sum_{i=1}^n \Phi_\pi(\hat{\mathcal{P}}_n)(O_i)$, with Φ_π as in Proposition 1 and with nuisance parameters replaced by their estimators $\hat{e}, \hat{q}_W, \hat{q}_L$ and $\hat{p}_W, \hat{p}_L$ (Equation (18) in Appendix A.5). The 1-step optimal writes as:

$$\hat{\pi}_{\text{1-step}} \in \operatorname*{argmax}_{\pi\in\Pi} \left\{\hat{V}_{\text{1-step}}(\pi) \overset{\text{def}}{=} \Psi^\pi_{\text{1-step}}\right\} . \quad (13)$$

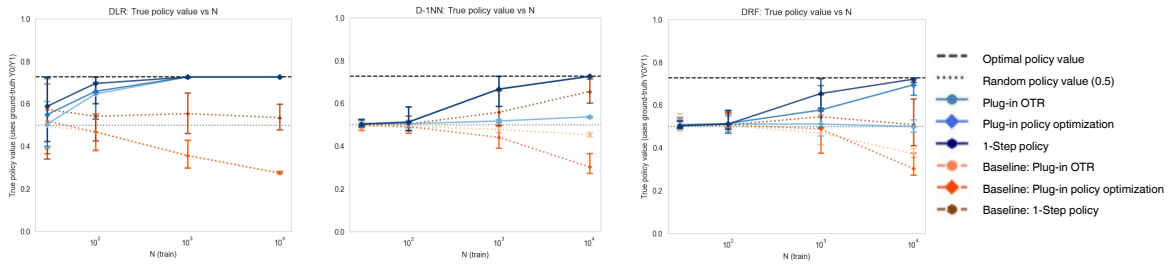
This policy inherits the properties of the 1-Step correction for policy values and is expected to perform better than plug-in OTR (Section 5.2) and plug-in value optimization policy (Section 5.3). In practice, the policy $\hat{\pi}_{\text{1-step}}$ is computed as in Section 5.1, using cross-fitted policy trees or weighted classification (Athey and Wager, 2021; Chernozhukov et al., 2018). However, policy value $\hat{V}_\pi$ is replaced by its debiased value $\psi^\pi_{\text{1-step}}$. In our experiments, the 1-Step correction consistently improves policy performance when combined with a plug-in value optimization policy.

7 EXPERIMENTS

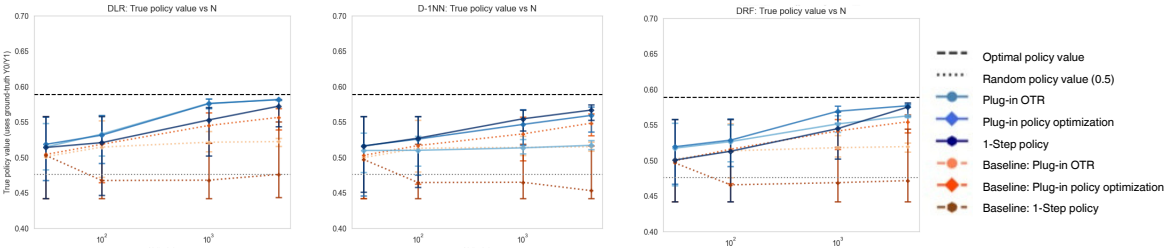
We conduct synthetic and semi-synthetic experiments to evaluate the performance of preference-based policy learning under our proposed approach, which estimates CPTE-related policy values. Our results demonstrate that the method consistently learns near-optimal policies even when key assumptions do not hold. We further benchmark its performance against both non-preference-based methods and policies with oracle access, highlighting the robustness and competitiveness of our approach.

7.1 Synthetic experiments

We consider a synthetic experimental setup with real-valued outcomes and $w(y|y') = \mathbb{1}_{\{y>y'\}}$, corresponding



(a) Synthetic experiment. Policy Learning with PNS policy value, heterogeneous RCT without correlation.



(b) Semi-synthetic experiment. Policy learning with hierarchical outcomes, no potential outcomes correlation.

Figure 1: Policy learning results on synthetic and semi-synthetic experiments with no outcome correlations

to the PNS application. The data-generating process is detailed in Appendix E.1.1 and satisfies Assumptions 1 to 3. Our synthetic experimental setup explores the benefits of preference-based policy learning with CPTE, contrasted with classical CATE policies. Outcomes are generated so that each unit's PNS prefers one treatment, while the CATE indicates the other one. This can occur when one of the potential outcomes distributions has a heavy tail distribution. Figures 2a and 2b (Appendix E) illustrate two examples of such potential outcome distributions. Our data-generating process compares the following properties and their impact on preference-based policy learning: *(i)* RCT vs. observational study; *(ii)* Homogeneous vs. heterogeneous treatment effect (Definition 3); *(iii)* Correlation between counterfactual outcomes (negative, non-correlated, or positive correlation), violating the identifiability conditions of the Conditional ITE in Theorem 1.

We examine policy learning and evaluation for different combinations of the data generating parameters. We show results of the three different methods introduced in this paper to sample from the conditional outcome distribution $P(Y|X,T)$ and estimate the CPTE: distributional $k-$NN with varying values of k as in Section 4.1 (with $k = 1$ in Figure 1), distributional random forests (DRF) and distributional linear regression (DLR) (Section 4.2). We compare our CPTE policies to CATE policies learned using a T-learner framework and corresponding base-learners: standard $k-$NN, ridge regression and regression forests (Künzel et al., 2019), which all estimate $\mathbb{E}[Y|X,T]$ instead of sampling from $P(Y|X,T)$ like the distributional methods. These are denoted as *baselines* in our figures. The policy is then learned using the following approaches: plug-in OTR (Section 5.2); plug-in policy value optimization (Section 5.3); and policy value optimization with the EIF-corrected value (1-Step corrected policy, Section 6). Policies are trained 50 times, and evaluated on a held-out set with oracle knowledge.

In Figure 1a, we present the heterogeneous RCT scenario with uncorrelated potential outcomes. The figure shows the preference policy value with 95% bootstrap confidence intervals. The EIF correction consistently improves policy learning. Because the data-generating process is linear, DLR performs best, while DRF and KNN converge more slowly. Results for policy value estimation are provided in fig. 4, Appendix E.1.3. Plug-in and EIF policy evaluation converge to the true policy value, while the baseline CATE-based policies diverge to 0 or 1. Policy learning results for other combinations of the data generating process (RCT/Observational, homogeneous/heterogeneous, correlated potential outcomes) are deferred to Figure 3, Appendix E.1.3. Importantly, while correlated potential outcomes significantly bias policy value estimation, the learned policy itself still obtains good oracle performances. Even when the identifiability assumptions of Theorem 1 do not hold, our methodology still learns good preference-based policies.

7.2 Semi-synthetic experiments

We conduct a semi-synthetic experiment inspired by the Tennessee STAR study (Word et al., 1990), a landmark randomized controlled trial on the effects of class size on student achievement. The study followed over 11,000 students from kindergarten through third grade across 79 schools in Tennessee, randomly assigning them to small, regular-sized or regular classes with a teacher's aide. Student outcomes were measured in core subjects, including math and reading, along with detailed records of participation and attrition. In our setup, we treat math scores as a secondary continuous outcome and construct a binary primary outcome from student attrition, using the same preference function as in Equation (1) and (Pocock et al., 2011). We collapse regular classrooms with and without aides, yielding 1,762 small-class students and 4,109 regular-class students after filtering. We generate synthetic outcomes for this real data by training an MLP and using MC-dropout (Gal and Ghahramani, 2015), approximating conditional distributions. These then serve as ground-truth for optimal policy and value. The full data generating process is detailed in Appendix E.2.1.

We compare the same estimation and policy learning methods as in the synthetic experiments. The semi-synthetic setting allows us to control correlations between potential outcomes, while still being constrained by real-world covariates and approximating observed outcome distributions. We provide experiments on non-correlated potential outcomes, negative correlations and positive correlations, using the Iman-Conover method (Iman and Conover, 1982). Results on the semi-synthetic STAR data show that 'preference'-based policies outperform mean-based ones, as they balance hierarchical outcomes, prioritizing attrition while leveraging math scores. Unlike our synthetic experiments, the preference functions are close to the decision boundary, making policy learning more difficult. This is highlighted by Figure 1b. DLR is misspecified, EIF corrects DRF and KNN for larger samples. Baseline methods perform poorly. Similarly to the synthetic experiments, adding strong correlations between potential outcomes does not impact policy learning performance. Figures 8 and 11 (Appendix E.2) respectively summarizes all semi-synthetic policy learning results and shows policy agreement results relative to the hierarchical oracle policy, and oracle policies that solely optimize the primary or secondary outcome.

8 CONCLUSION

This paper provides a general methodology for policy learning and for quantifying conditional treatment effects with complex outcomes that are best assessed via comparison. We define *(a)* an unidentifiable conditional treatment effect (Equation (2)); *(b)* an identifiable one (Definition 2); *(c)* and offer estimators for the identifiable quantity (Section 4). Our experiments show that our estimators and policy learning methods efficiently learn preference-based policies, linking *(b)* and *(c)*. They also showcase when preference-based policies may be preferable to CATE-based policies. Lastly, we explore the effects of violating the identifiability assumptions connecting b to a by adding correlation to the potential outcomes in both the synthetic (appendix E.1.3) and semi-synthetic (appendix E.2.2) experiments. We find policy evaluation to be sensitive to this violation, while policy learning was negligibly impacted.

Our work paves the way to several orthogonal research questions for future works, including performing sensitivity analyses to assess when the conditional measure defined in Definition 2 is close to the non-identifiable one, and leveraging our framework with unknown preference functions w that would need to be learned while optimizing policies.

Acknowledgments

We thank the anonymous reviewers for their helpful feedback which helped improve the quality of our paper. M. Even acknowledges funding from Theremia. D.Parnas and U.Shalit were supported by ISF grant 2456/23.

References

Alberto Abadie, Joshua Angrist, and Guido Imbens. Instrumental variables estimates of the effect of subsidized training on the quantiles of trainee earnings. *Econometrica*, 70(1):91–117, 2002. doi: 10.1111/1468-0262.00270. 2

C.M. Achilles, Helen Pate Bain, Fred Bellott, Jayne Boyd-Zaharias, Jeremy Finn, John Folger, John Johnston, and Elizabeth Word. Tennessee's Student Teacher Achievement Ratio (STAR) project, 2008. URL https://doi.org/10.7910/DVN/SIWH9F. 27, 28

Susan Athey and Guido Imbens. Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences*, 113(27):7353–7360, 2016. 2

Susan Athey and Stefan Wager. Policy learning with observational data. *Econometrica*, 89(1):133–161, 2021. 2, 3, 6, 7

Gérard Biau and Luc Devroye. *Lectures on the nearest neighbor method*, volume 246. Springer, 2015. 2

Marc Buyse. Generalized pairwise comparisons of prioritized outcomes in the two-sample problem. *Statistics in Medicine*, 29(30):3245–3257, December 2010. ISSN 1097-0258. doi: 10.1002/sim.3923. URL http://dx.doi.org/10.1002/sim.3923. 2, 6

Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. Double/debiased machine learning for treatment and structural parameters, 2018. 7

Bénédicte Colnet, Julie Josse, Gaël Varoquaux, and Erwan Scornet. Causal effect on a target population: A sensitivity analysis to handle missing covariates. *Journal of Causal Inference*, 10(1):372–414, 2022. 4

Bénédicte Colnet, Julie Josse, Gaël Varoquaux, and Erwan Scornet. Risk ratio, odds ratio, risk difference... which causal measure is easier to generalize? *arXiv preprint arXiv:2303.16008*, 2023. 4

Yifan Cui and Sukjin Han. Policy learning with distributional welfare. *arXiv preprint arXiv:2311.15878*, 2023. 2, 3

Johan de Aguas, Sebastian Krumscheid, Johan Pensar, and Guido Biele. The probability of tiered benefit: Partial identification with robust and stable inference. *arXiv preprint arXiv:2502.10049*, 2025. 2

Gaohong Dong, Bo Huang, Johan Verbeeck, Ying Cui, James Song, Margaret Gamalo-Siebers, Duolao Wang, David C Hoaglin, Yodit Seifu, Tobias Mütze, et al. Win statistics (win ratio, win odds, and net benefit) can complement one another to show the strength of the treatment effect on time-to-event outcomes. *Pharmaceutical Statistics*, 22(1):20–33, 2023. 27

Mathieu Even and Julie Josse. Rethinking the win ratio: A causal framework for hierarchical outcome analysis. *arXiv preprint arXiv:2501.16933*, 2025. 1, 2, 4

Douglas Faries, Xiang Zhang, Zbigniew Kadziola, Uwe Siebert, Felicitas Kuehne, Robert L Obenchain, and Josep Maria Haro. *Real world health care data analysis: causal methods and implementation using SAS*. SAS Institute, 2020. 29

Sergio Firpo. Efficient semiparametric estimation of quantile treatment effects. *Econometrica*, 75 (1):259–276, 2007. doi: 10.1111/j.1468-0262.2007.00735.x. 2

Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation. *arXiv preprint arXiv:1506.02157*, 2015. 9, 27, 28

William H Greene. Econometric analysis. *Pretence Hall*, 2003. 2, 6

Oliver Hines, Oliver Dukes, Karla Diaz-Ordaz, and Stijn Vansteelandt. Demystifying statistical learning based on efficient influence functions. *The American Statistician*, 76(3):292–304, 2022. 7, 15

Paul W. Holland. Statistics and causal inference. *Journal of the American Statistical Association*, 81(396): 945–960, 1986. ISSN 01621459, 1537274X. URL http://www.jstor.org/stable/2289064. 2, 23

Chia-Ling Huang, John Lamb, Leonid Chindelevitch, Jarek Kostrowicki, Justin Guinney, Charles DeLisi, and Daniel Ziemek. Correlation set analysis: detecting active regulators in disease populations using prior causal knowledge. *BMC bioinformatics*, 13 (1):46, 2012. 29

Emily J Huang, Ethan X Fang, Daniel F Hanley, and Michael Rosenblum. Inequality in treatment benefits: Can we determine if a new treatment benefits the many or the few? *Biostatistics*, 18(2):308–324, 2017. 2

Ronald L Iman and William-Jay Conover. A distribution-free approach to inducing rank correlation among input variables. *Communications in Statistics-Simulation and computation*, 11(3):311–334, 1982. 9, 29

Ying Jin, Zhimei Ren, and Emmanuel J Candès. Sensitivity analysis of individual treatment effects: A robust conformal inference approach. *Proceedings of the National Academy of Sciences*, 120(6): e2214889120, 2023. 4

Nathan Kallus and Miruna Oprescu. Robust and agnostic learning of conditional distributional treatment effects. In *International Conference on Artificial Intelligence and Statistics*, pages 6037–6060. PMLR, 2023. 2

Yuta Kawakami, Manabu Kuroki, and Jin Tian. Probabilities of causation for continuous and vector variables. *arXiv preprint arXiv:2405.20487*, 2024. 2

Christian Kleiber and Achim Zeileis. *Applied Econometrics with R*. Springer-Verlag, New York, 2008. doi: 10.1007/978-0-387-77318-6. URL https://CRAN.R-project.org/package=AER. 28

Roger Koenker and Gilbert Bassett. Regression quantiles. *Econometrica*, 46(1):33, January 1978a. ISSN 0012-9682. doi: 10.2307/1913643. URL http://dx.doi.org/10.2307/1913643. 2, 6

Roger Koenker and Gilbert Bassett. Regression quantiles. *Econometrica*, 46(1):33–50, 1978b. doi: 10.2307/1913643. 2

Alan B Krueger. Experimental estimates of education production functions. *The quarterly journal of economics*, 114(2):497–532, 1999. 27, 28

Sören R Künzel, Jasjeet S Sekhon, Peter J Bickel, and Bin Yu. Metalearners for estimating heterogeneous treatment effects using machine learning. *Proceedings of the national academy of sciences*, 116(10): 4156–4165, 2019. 2, 3, 8

Lihua Lei. Causal interpretation of regressions with ranks. *arXiv preprint arXiv:2406.05548*, 2024. 2

Ang Li and Judea Pearl. Probabilities of causation with nonbinary treatment and effect. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pages 20465–20472, 2024. 2, 4

Jiannan Lu, Peng Ding, and Tirthankar Dasgupta. Treatment effects on ordinal outcomes: Causal estimands and sharp bounds. *Journal of Educational and Behavioral Statistics*, 43(5):540–567, 2018. 2

Charles F Manski. *Partial identification of probability distributions*. Springer, 2003. 2

Lu Mao. On causal estimation using u-statistics. *Biometrika*, 105(1):215–220, dec 2017. ISSN 1464-3510. 5

Nicolai Meinshausen and Greg Ridgeway. Quantile regression forests. *Journal of machine learning research*, 7(6), 2006. 2, 6

Scott Mueller and Judea Pearl. Personalized decision making–a conceptual introduction. *Journal of Causal Inference*, 11(1):20220050, 2023. 1, 2, 3

Jeffrey Näf and Herbert Susmann. Causal-drf: Conditional kernel treatment effect estimation using distributional random forest. *arXiv preprint arXiv:2411.08778*, 2024. 2

Trang Quynh Nguyen, Benjamin Ackerman, Ian Schmid, Stephen R Cole, and Elizabeth A Stuart. Sensitivity analyses for effect modifiers not observed in the target population when generalizing treatment effects from a randomized controlled trial: Assumptions, models, effect scales, data scenarios, and implementation details. *PloS one*, 13(12):e0208795, 2018. 4

Xinkun Nie and Stefan Wager. Quasi-oracle estimation of heterogeneous treatment effects. *Biometrika*, 108 (2):299–319, 2021. 3

Junhyung Park, Uri Shalit, Bernhard Schölkopf, and Krikamol Muandet. Conditional distributional treatment effect with kernel conditional mean embeddings and u-statistic regression. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8401–8412. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/park21c.html. 2

Judea Pearl. *Causality*. Cambridge University Press, Cambridge, England, September 2009. 4

François Petit, Gérard Biau, and Raphaël Porcher. Optimal treatment regimes for the net benefit of a treatment. *arXiv preprint arXiv:2503.22580*, 2025. 2, 4

S. J. Pocock, C. A. Ariti, T. J. Collier, and D. Wang. The win ratio: a new approach to the analysis of composite endpoints in clinical trials based on clinical priorities. *European Heart Journal*, 33(2): 176–182, September 2011. ISSN 1522-9645. doi: 10.1093/eurheartj/ehr352. URL http://dx.doi.org/10.1093/eurheartj/ehr352. 1, 2, 3, 9

James M Robins and Thomas S Richardson. Alternative graphical causal models and the identification of direct effects. *Causality and psychopathology: Finding the determinants of disorders and their cures*, 84:103–158, 2010. 2

Murray Rosenblatt. Remarks on a multivariate transformation. *The annals of mathematical statistics*, 23 (3):470–472, 1952. 6

Skipper Seabold and Josef Perktold. statsmodels: Econometric and statistical modeling with python. In *9th Python in Science Conference*, 2010. 6

Ryusei Shingaki et al. Identification and estimation of joint probabilities of potential outcomes in observational studies with covariate information. *Advances in neural information processing systems*, 34:26475–26486, 2021. 2

Jerzy Splawa-Neyman, Dorota M Dabrowska, and Terrence P Speed. On the application of probability theory to agricultural experiments. essay on principles. section 9. *Statistical Science*, pages 465–472, 1990. 3

Erik Sverdrup, Ayush Kanodia, Zhengyuan Zhou, Susan Athey, and Stefan Wager. policytree: Policy learning via doubly robust empirical welfare maximization over trees. *Journal of Open Source Software*, 5(50):2232, 2020. 7

Xiaoqing Tan, Shu Yang, Wenyu Ye, Douglas E Faries, Ilya Lipkovich, and Zbigniew Kadziola. Double machine learning methods for estimating average treatment effects: a comparative study. 2025. 29

Olivier Thas, Jan De Neve, Lieven Clement, and Jean-Pierre Ottoy. Probabilistic index models. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 74(4):623–671, 2012. 2, 4

Jin Tian and Judea Pearl. Probabilities of causation: Bounds and identification. *Annals of Mathematics and Artificial Intelligence*, 28(1):287–313, 2000. 1, 2, 5

Aad W Van der Vaart. *Asymptotic statistics*, volume 3. Cambridge university press, 2000. 18

Stefan Wager and Susan Athey. Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 113(523):1228–1242, 2018. 3

Stefan Wager, Susan Athey, and Ryan Tibshirani. Generalized random forests. *Annals of Statistics*, 47(2):1148–1178, 2019. 2

Shuai Wang and Ang Li. Estimating probabilities of causation with machine learning models. *arXiv preprint arXiv:2502.08858*, 2025. 2

Shuai Wang, Song Jiang, Yizhou Sun, Judea Pearl, and Ang Li. Learning probabilities of causation from finite population data. *arXiv preprint arXiv:2505.17133*, 2025. 2

Elizabeth Word, John Johnston, Helen P Bain, B DeWayne Fulton, Jayne B Zaharias, Charles M Achilles, Martha N Lintz, John Folger, and Carolyn Breda. The state of tennessee's student/teacher achievement ratio (star) project. *Tennessee Board of Education*, 1990. 9, 27

Tianhui Zhou, William E Carson IV, and David Carlson. Estimating potential outcome distributions with collaborating causal networks. *Transactions on machine learning research*, 2022:https–openreview, 2022. 2

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. Yes.
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. No.
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. No.

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. Yes.
   (b) Complete proofs of all theoretical results. Yes.
   (c) Clear explanations of any assumptions. Yes.

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). Yes.
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). Yes.
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). Yes.
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). No.

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. Yes.
   (b) The license information of the assets, if applicable. No.
   (c) New assets either in the supplemental material or as a URL, if applicable. No.
   (d) Information about consent from data providers/curators. Not Applicable.
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. Not Applicable.

5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. Not Applicable.
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. Not Applicable.
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. Not Applicable.

A FURTHER SPECIFIC DETAILS

A.1 Illustrative Adversarial Examples

Example 1: observed and structural treatment effect modifiers. Let us place ourselves in a simple setting where we assume that features $X_i \in \mathcal{X}$ *do not* include treatment effect modifiers, denoted as $O_i \in \mathcal{O}$. Assume that $\mathcal{Y} \subset \mathbb{R}$, $w(y|y') = \mathbb{1}_{\{y>y'\}}$, the set of treatment effect modifiers $\mathcal{O}$ is of the form $\mathcal{O} = \mathcal{O}_1 \cup \mathcal{O}_2$ with $\mathbb{P}(O_i \in \mathcal{O}_1) = 1-\alpha$, $\mathbb{P}(O_i \in \mathcal{O}_2) = \alpha$ and O_i is independent of X_i, such that:

$$Y_i(1) = y_1 > y_0 = Y_i(0)|O_i \in \mathcal{O}_1\,,$$
$$Y_i(1) = y_1' < y_0' = Y_i(0)|O_i \in \mathcal{O}_2\,,$$

almost surely. We then have, if $y_0' > y_1 > y_0 > y_1'$:

$$\mathbb{E}\left[w(Y(1,X_i)|Y_i(0))|X_i\right] = (1-\alpha)^2\,, \tag{14}$$
$$\mathbb{E}\left[w(Y_i(1)|Y_i(0))|X_i\right] = 1-\alpha\,, \tag{15}$$
$$\mathbb{E}\left[w(Y(1,X_i)|Y_i(0))|X_i, O_i\right] = \mathbb{1}_{\{O_i \in \mathcal{O}_1\}}\,. \tag{16}$$

Equation (14) is the CPTE when we observe only X_i, which is non-informative. If we observe also the informative variable O_i, in Equation (16), heterogeneity is fully captured with the treatment effect modifier O_i. However, Equation (15) can also be written as the average of Equation (16) conditionally on X_i: but we don't obtain the same value as in Equation (14). This is due to the fact that O_i is also a *structural* treatment effect modifier.

Example 2: linear CATE *vs* CPTE. We now present an example showing different treatment recommendations between a classical linear CATE and our preference-based conditional effects. The example illustrates how CPTE leverages more than first-order moment information, and follows a majority-vote approach for treatment recommendations. Let $\mathcal{Y} = \mathbb{R}$ and

$$Y_i(1) \sim \frac{\delta_0 + \delta_{0.1} + \delta_{1.1}}{3}|X_i = x\,,$$
$$Y_i(0) \sim \frac{\delta_{-0.1} + \delta_1}{2}|X_i = x\,.$$

In this case, $\text{CPTE}(x) = 0.66 > 0.5$, while $\mathbb{E}\left[Y_i(1) - Y_i(0)|x\right] = 0.4 - 0.45 = -0.05 < 0$. Treatment recommendations differ depending on the policy maker's objective - whether they consider the magnitude of the effects (risk difference) or the most likely treatment to benefit unit X_i. This archetype is the basis of our synthetic experiments.

A.2 Translation to Structural Causal Model (SCM) terminology.

Here we reformulate several of our basic definitions and theorem's within Pearl's Structural Causal Model (SCM) framework. Let $(\mathcal{G}, U, f)$ be an SCM, where U denotes exogenous variables, $V = (T, Y, X)$ endogenous variables, and

$$Y = f_Y(T, X, U_Y).$$

Counterfactual outcomes are defined by interventions as

$$Y(t) := Y \mid \text{do}(T = t).$$

Preference-based individual causal effect. For a unit characterized by exogenous variables U, the individual-level preference-based causal effect is

$$\Delta(U_i) := w\big(f_Y(1, X_i, U_i),\ f_Y(0, X, U_i)\big)\,.$$

Note that we have indeed $\Delta_i = \Delta(U_i)$. The conditional ITE is then equal to:

$$\mathbb{E}\left[w\big(f_Y(1, X_i, U_i),\ f_Y(0, X, U_i)\big)\right]\,.$$

SCM interpretation of CPTE - Definition 2. For a given covariate value $X = x$, the Conditional Preference Treatment Effect CPTE(x) corresponds to comparing outcomes generated under independent draws of the exogenous variables:

$$U \sim P(U \mid X = x), \qquad U' \sim P(U \mid X = x), \qquad U \perp U',$$

and $Y(1, x)$, $Y(0, x)$ defined in Definition 2 can also be defined as:

$$Y(1, x) = f_Y(1, x, U_Y), \qquad Y(0, x) = f_Y(0, x, U'_Y),$$

yielding the same CPTE definition:

$$\text{CPTE}(x) = \mathbb{E}\big[w\big(Y(1, x), Y_i(0)\big)\big] = \mathbb{E}\big[w\big(Y_i(1), Y(0, x)\big)\big]$$

Structural treatment effect modifiers (SCM) - Definition 4. A set of variables $Z_I \subseteq U$ is a set of *structural treatment effect modifiers* if for all $i, j \in [n]$ and all $x \in \mathbb{R}^{\mathcal{I}}$:

$$\mathbb{E}\left[w(Y_i(1), Y_j(0)) \mid Z_i, Z_j, (Z_i^{\mathcal{I}}, Z_j^{\mathcal{I}}) = (x, x)\right] = g(x)\,,$$

where $z^{\mathcal{I}} = (z_k)_{k \in \mathcal{I}}$ for $z \in \mathbb{R}^p$.

Conditions for CPTE equivalence to individual-level estimation - Theorem 1.
Assume that structural treatment effect modifiers in the sense of Definition 4 are included in the set of features X_i. Then, CPTE$(x) = \mathbb{E}\left[w(Y(1, U_i)|Y(0, U_i))|X_i = x\right]$ for all $x \in \mathcal{X}$.

A.3 Correcting the plug-in bias with efficient influence functions

Plug-in estimators for the nuisance functions in Equation (11) can cause large bias to $\hat{V}(\pi)$ if the models have large non-asymptotic biases (e.g. nearest-neighbors), if they overfit (e.g. random forests) or if they are misspecified (e.g. linear models). Efficient Influence Functions (EIFs) can be used to correct for this plug-in bias (Hines et al., 2022) and to obtain efficient estimators that satisfy, under mild assumptions, double robustness properties, semi-parametric efficiency optimality and minimal asymptotic variance among certain classes of estimators.

Let $\Psi(\mathcal{P})$ be an estimand we wish to estimate, that is the function of an observed distribution $\mathcal{P}$. In our case, $\mathcal{P}$ is the distribution of our n observations $O_i = (X_i, T_i, Y_i)$, and the estimand is $\Psi_\pi(\mathcal{P}) = V(\pi)$, so that:

$$\Psi_\pi(\mathcal{P}) = \mathbb{E}\left[\pi(X_i) q_W(X_i) + (1 - \pi(X_i) q_L(X_i))\right]\,.$$

As we saw before, in order to estimate $\Psi_\pi(\mathcal{P})$, we first need (non-parametric) nuisance estimates (of q_W and q_L) learned on the data. However, small perturbations and errors in these nuisance estimations can lead to biases in our estimate of Ψ_π if we simply plug-in the nuisance estimators like in Equation (6). This is the *plug-in policy estimator* of Ψ_π, denoted as $\Psi^\pi_{\text{plug-in}}$.

Efficient Influence Function (EIF). The Efficient Influence Function $\phi_\pi(\mathcal{P})$ of Ψ_π at $\mathcal{D}$ for any observation o can be written:

$$\phi(\mathcal{P})(o) \stackrel{\text{def}}{=} \lim_{t \to 0} \frac{\mathrm{d}\psi(\mathcal{P}_t)}{\mathrm{d}t}\,, \quad \text{where} \quad \mathcal{P}_t = (1 - t)\mathcal{P} + t\delta_o\,,$$

where δ_o is a Dirac at o. It should be understood as a 'gradient' of Ψ_π, computed at $\mathcal{P}$ and evaluated on observation o. Hines et al. (2022) provides methodologies to compute these EIFs with examples on classical estimands such as the ATE, ATT or RR. The 1-step estimator corrects first-order biases that come from wrongly estimating the nuisance parameters of the distribution, where $\Psi^\pi_{\text{plug-in}} = \hat{V}(\pi)$ (Equation (11)) and Φ^π is its EIF:

$$\Psi^\pi_{\text{1-step}} \stackrel{\text{def}}{=} \Psi^\pi_{\text{plug-in}} + \frac{1}{n} \sum_{i=1}^{n} \Phi^\pi(\hat{\mathcal{P}}_n)(O_i)\,. \tag{17}$$

A.4 Distributional regression algorithm

The meta-algorithm Algorithm 1 outputs estimators $\hat{q}_W, \hat{q}_L : \mathcal{X} \mapsto [0, 1]$ of $q_W(x) = \mathbb{E}\left[w(Y_i(1), Y(0, X_i))|X_i = x\right] = \text{CPTE}(x)$ and $q_L(x) = \mathbb{E}\left[w(Y_i(0), Y(1, X_i))|X_i = x\right]$ (the CPTE with reversed treatments) from a given quantile models $M_q^{(t)}(x)$ and an interpolation rule. First, Algorithm 1

discretizes $[0,1]$ into $\mathcal{Q}$, and computes values $Y_q(t,x)$ for counterfactuals corresponding to these quantiles. This step estimates the inverse CDF. Then S other quantiles are sampled, and values corresponding to these quantiles are computed using the interpolation rule, sampling from the inverse CDF. `Interp` is any function that takes as inputs a set $(q, f(q))_{q\in\mathcal{Q}}$ and a value $x \in [0,1]$ where $\mathcal{Q}$ is a discretization of $[0,1]$, $f : [0,1] \to \mathcal{Y}$ is a model evaluated only on discretized values. `Interp` then outputs a value for $f(x)$. The obtained values are then averaged in a final step.

Algorithm 1 Sampling-based Estimation of q_W, q_L via Quantile Models

Require: Quantile models $M_q^{(1)}(x)$ and $M_q^{(0)}(x)$; grid size $|\mathcal{Q}|$; number of samples S; interpolation rule `Interp`.
Ensure: Estimator q_W, q_L.

1: Sample a grid $\mathcal{Q} = \{q_1, \dots, q_{|\mathcal{Q}|}\}$ with $q_k \sim \mathrm{Unif}(0,1)$.
2: For each $q \in \mathcal{Q}$, compute

$$Y_q(1,x) \leftarrow M_q^{(1)}(x), \qquad Y_q(0,x) \leftarrow M_q^{(0)}(x).$$

3: Draw S independent quantiles $u_s^{(1)}, u_s^{(0)} \sim \mathrm{Unif}(0,1)$ for $s = 1, \dots, S$.
4: For each s, interpolate:

$$Y_s(1,x) \leftarrow \texttt{Interp}\big(\{(q, Y_q(1,x))\}_{q\in\mathcal{Q}},\, u_s^{(1)}\big),$$

$$Y_s(0,x) \leftarrow \texttt{Interp}\big(\{(q, Y_q(0,x))\}_{q\in\mathcal{Q}},\, u_s^{(0)}\big).$$

5: Estimate

$$\widehat{q}_W(x) \leftarrow \frac{1}{S}\sum_{s=1}^{S} w\big(Y_s(1,x),\, Y_s(0,x)\big).$$

$$\widehat{q}_L(x) \leftarrow \frac{1}{S}\sum_{s=1}^{S} w\big(Y_s(0,x),\, Y_s(1,x)\big).$$

A.5 The 1-Step policy value estimator

The 1-step policy value estimator, defined in Equation (17), writes as:

$$\begin{aligned}\Psi^{\pi}_{1-\text{step}} = \frac{1}{n}\sum_{i=1}^{n}\Big[&\pi(X_i)\hat{q}_W(X_i) + (1-\pi(X_i))\hat{q}_L(X_i)\\ &+ \frac{T_i}{e(X_i)}\Big(\pi(X_i)(\hat{p}_W(X_i,T_i,Y_i) - \hat{q}_W(X_i))\\ &+ (1-\pi(X_i))(\hat{p}_L(X_i,T_i,Y_i) - \hat{q}_L(X_i))\Big)\\ &+ \frac{1-T_i}{1-e(X_i)}\Big(\pi(X_i)(\hat{p}_W(X_i,T_i,Y_i) - \hat{q}_W(X_i))\\ &+ (1-\pi(X_i))(\hat{p}_L(X_i,T_i,Y_i) - \hat{q}_L(X_i))\Big)\Big].\end{aligned} \tag{18}$$

B MISCELLANEOUS PROOFS

B.1 Proof of Lemma 1

Proof. Assume that $\mathcal{Y} = \{0,1\}$, such that $Y_i(1), Y_i(0)$ are Bernoulli variables of mean 0.5. We take $Y_i(1), Y_i(0)$ independent of X_i. Coupling $(Y_i(1), Y_i(0))|X_i = x$ as $Y_i(1) = Y_i(0)$ gives $\mathbb{P}(Y_i(1) > Y_i(0)) = 0$, while taking independent potential outcomes leads to $\mathbb{P}(Y_i(1) > Y_i(0)) = \mathbb{P}(Y_i(1) = 1, Y_i(0) = 0) = \frac{1}{4}$. Since the distribution of the observations (X_i, T_i, Y_i) does not change by taking either coupling but the value of $\mathbb{E}[w(Y_i(1)|Y_i(0))]$ changes, we have non-identifiability: $\mathbb{E}[w(Y_i(1)|Y_i(0))]$ cannot be expressed as a function of the law of the observations. □

B.2 Proof of Theorem 1

Proof. If $X_i = Z_i^{\mathcal{I}}$ with $\mathcal{I}$ the set of structural treatment effect modifiers or if $Z_i \subset X_i$, there exists g such that $\mathbb{E}\left[w(Y_i(1)|Y_j(0))|Z_i, Z_j, (X_i, X_j) = x\right] = g(x)$. Then,

$$\begin{aligned}\mathbb{E}\left[w(Y_i(1)|Y_j(0)|X_i = X_j = x\right] &= \mathbb{E}\left[\mathbb{E}\left[w(Y_i(1)|Y_j(0)|Z_i, Z_j, (Z_i^{\mathcal{I}}, Z_j^{\mathcal{I}}) = (x, x)\right]|X_i = X_j = x\right] \\ &= \mathbb{E}\left[g(x)|X_i = X_j = x\right] \\ &= g(x)\,.\end{aligned}$$

Now, taking $i = j$,

$$\begin{aligned}\mathbb{E}\left[w(Y_i(1)|Y_i(0)|X_i = x\right] &= \mathbb{E}\left[\mathbb{E}\left[w(Y_i(1)|Y_i(0)|Z_i, Z_i^{\mathcal{I}} = x\right]|X_i = x\right] \\ &= \mathbb{E}\left[g(x)|X_i = x\right] \\ &= g(x)\,,\end{aligned}$$

concluding the proof. □

C PROOF OF Theorem 2

Let $\mathbf{X_t} = \{X_i, T_i = t\}$ for arms $t = 0, 1$ and $\mathbf{X} = \mathbf{X}_0 \cup \mathbf{X}_1$. For any $x \in \mathcal{X}$, let $Y(1, x)$ and $Y(0, x)$ be independent copies of $Y_i(1)|X_i = x$ and $Y_i(0)|X_i = x$ respectively, which are also independent of each other. $Y(1, x)$ and $Y(0, x)$ are independent.
Let the $k-$nearest neighborhood of arm $t \in \{0, 1\}$ of some point $x \in \mathcal{X}$ be:

$$\mathcal{N}_t(x, k) = \{a \in \mathbf{X_t}\ s.t.\ |\{b \in \mathbf{X_t} : ||b - x|| < ||a - z||\}| < k\}$$

A generic neighborhood, independent of treatment assignment, is designated $\mathcal{N}(x, k)$. First,

$$\begin{aligned}\mathbb{E}\left[\hat{q}(x)\right] &= \mathbb{E}\left[w(Y_i|Y_j)|(i, j) \in \mathcal{N}_1(x, k) \times \mathcal{N}_0(x, k)\right] \\ &= \mathbb{E}\left[w(Y_i(1)|Y_j(0))|(i, j) \in \mathcal{N}_1(x, k) \times \mathcal{N}_0(x, k)\right] \quad \text{(Assumptions 1 and 2)}\,.\end{aligned}$$

Step 1: $\mathbb{E}\left[\hat{q}(x)\right] \to q(x)$**.** We have:

$$\begin{aligned}\mathbb{E}\left[\hat{q}(x)\right] - q(x) &= \mathbb{E}\left[w(Y_i(1)|Y_j(0))|(i, j) \in \mathcal{N}_1(x, k) \times \mathcal{N}_0(x, k)\right] - \mathbb{E}\left[w(Y(1, x), Y(0, x))\right] \\ &= \mathbb{E}\left[w(Y_i(1)|Y_j(0)) - w(Y(1, x), Y(0, x))|(i, j) \in \mathcal{N}_1(x, k) \times \mathcal{N}_0(x, k)\right]\,.\end{aligned}$$

We then introduce the following classical lemma, which we prove for completeness.

Lemma 2. *Let $x \in \mathcal{X}$ and any $\hat{x} \in \mathcal{N}(x, k)$. We have:*

$$\hat{x} \xrightarrow{p} x$$

when $k = o(n/log(n))$.

Proof. We need to show:

$$\forall \epsilon > 0\,, \quad \lim_{n\to\infty} P(||x - \hat{x}|| > \epsilon)\,.$$

Let $B(x, \epsilon)$ be a sphere of ϵ radius around x and let $p = \mathbb{P}(\hat{x} \in B(x, \epsilon))$ be the probability of $\hat{x}$ being in the ϵ-sphere. We denote $Z = |\{z \in \mathbf{X}|z \in B(x, \epsilon)\}|$, the number of datum that fall within the ϵ-sphere. We note that $Z \sim Bin(N, p)$, leading to:

$$\lim_{n\to\infty} P(||x - \hat{x}|| > \epsilon) \leqslant \lim_{n\to\infty} \mathbb{P}(Z < k) = \lim_{n\to\infty} \sum_{i=0}^{k-1} \binom{n}{i} p^i (1 - p)^{n-i}$$

If $p \in (0, 1)$,

$$\lim_{n\to\infty} \sum_{i=0}^{k-1} \binom{n}{i} p^i (1 - p)^{n-i} \leqslant \lim_{n\to\infty} kn^k q^n$$

where $q = \max(p, 1-p) < 1$ and where we bounded $\binom{n}{i}$ by $n^i \leqslant n^k$. Thus, since $kn^k q^n = \exp(\log(k) + l\log(n) + n\log(q))$ and $\log(q) < 0$, as long as $\frac{k\log(n)}{n} \to 0$:

$$\lim_{n\to\infty} P(||x - \hat{x}|| > \epsilon)\,,$$

hence the result. Then, if $p \notin (0,1)$, either $p = 0$ or $p = 1$. $p = 0$ corresponds to $x \notin \mathrm{Supp(X_i)}$ which we do not consider (since q will be evaluated on sampled datapoints) while if $p = 1$ it means that X_i is a deterministic random variable equal to x, meaning that the lemma is trivial. □

Since we assume that $w : (x, x') \to q(x, x')$ is continuous almost everywhere, we conclude using the continuous mapping theorem (Van der Vaart, 2000). Thus, $\mathbb{E}\left[\hat{q}(x)\right] \to q(x)$. To have consistency, note that

$$\mathbb{P}\left(|\hat{q}(x) - q(x)| > \epsilon\right) = \mathbb{P}\left(|\hat{q}(x) - \mathbb{E}[\hat{q}(x)] + \mathbb{E}[\hat{q}(x)] - q(x)| > \epsilon\right)\,.$$

Using an union bound,

$$\mathbb{P}\left(|\hat{q}(x) - q(x)| > \epsilon\right) \leqslant \mathbb{P}\left(|\hat{q}(x) - \mathbb{E}[\hat{q}(x)]| > \epsilon/2\right) + \mathbb{P}\left(|\mathbb{E}[\hat{q}(x)] - q(x)| > \epsilon/2\right)\,.$$

We thus need to show that:

$$\forall \varepsilon > 0\,, \quad \mathbb{P}\left(|\hat{q}(x) - \mathbb{E}[\hat{q}(x)]| > \epsilon\right) \to 0\,.$$

Step 2: $|\hat{q}(x) - \mathbb{E}[\hat{q}(x)]| \to 0$ **in probability.** We have:

$$\begin{aligned}\mathbb{P}\left(|\hat{q}(x) - \mathbb{E}[\hat{q}(x)]| > \epsilon\right) &= \mathbb{P}\left(\left|\frac{1}{k^2}\sum_{(i,j)\in\mathcal{N}_1(x,k)\times\mathcal{N}_0(x,k)} w(Y^i, Y^j) - \mathbb{E}[w(Y^i, Y^j)]\right| > \epsilon\right)\\ &\leqslant \frac{\mathbb{E}\left[\left(\sum_{(i,j)\in\mathcal{N}_1(x,k)\times\mathcal{N}_0(x,k)} w(Y^i, Y^j) - \mathbb{E}[w(Y^i, Y^j)]\right)^2\right]}{k^4\epsilon^2}\,,\end{aligned}$$

using Chebyshev's inequality. We will denote $d(i,j) := w(Y^i, Y^j) - \mathbb{E}[w(Y^i, Y^j)]$. We can rewrite the earlier expression as:

$$\frac{\mathbb{E}\left[\left(\sum_{(i,j)\in\mathcal{N}_1(x,k)\times\mathcal{N}_0(x,k)} d(i,j)\right)^2\right]}{k^4\epsilon^2}\,.$$

We split the quadratic expressions into two groups, depending on if there is overlap between the pairs compared in the preference function. We will denote the four-way cross product by $\tilde{\mathcal{N}} = \mathcal{N}_1(x,k) \times \mathcal{N}_1(x,k) \times \mathcal{N}_0(x,k) \times \mathcal{N}_0(x,k)$.

$$\frac{1}{k^4\epsilon^2}\sum_{(i,\hat{i},j,\hat{j})\in\tilde{N}} \left(\mathbb{I}\{i = \hat{i} \text{ or } j = \hat{j}\} + \mathbb{I}\{i \neq \hat{i}\}\mathbb{I}\{j \neq \hat{j}\}\right)\mathbb{E}[d(i,j)d(\hat{i},\hat{j})]$$

Case 1: There are $2k^3$ cases where at least $i = \hat{i}$ or $j = \hat{j}$.

$$\begin{aligned}&\mathbb{E}[d(i,j)d(\hat{i},\hat{j})]\\ &= \mathrm{cov}[w(Y^i, Y^j), w(Y^{\hat{i}}, Y^{\hat{j}})] \qquad \text{(definition of covariance)}\\ &\leqslant \mathrm{var}\left[w(Y^i, Y^j)\right]\\ &\leqslant 1\,,\end{aligned}$$

since w is assumed to be bounded by 1. Therefore, the summation may be bounded as follows:

$$\begin{aligned}\frac{1}{k^4\epsilon^2}\sum_{(i,\hat{i},j,\hat{j})\in\tilde{N}} \mathbb{I}\{i = \hat{i} \text{ or } j = \hat{j}\}\mathbb{E}[d(i,j), d(\hat{i},\hat{j})] &\leqslant \frac{2k^3}{k^4\epsilon^2}\\ &= \frac{2}{k\epsilon^2}\,.\end{aligned}$$

Case 2: There are $O(k^4)$ cases when $i \neq \hat{i}$ and $j \neq \hat{j}$. Because pairs are *i.i.d.*, if there are no shared members the covariance is zero. Therefore,

$$\frac{1}{k^4} \sum_{(i,\hat{i},j,\hat{j}) \in \tilde{N}} \mathbb{I}\{i \neq \hat{i}\}\mathbb{I}\{j \neq \hat{j}\}\mathbb{E}[d(i,j)d(\hat{i},\hat{j})] = \frac{1}{k^4} \sum_{(i,\hat{i},j,\hat{j}) \in \tilde{N}} \mathbb{I}\{i \neq \hat{i}\}\mathbb{I}\{j \neq \hat{j}\}\, Cov[w(Y^i, Y^j), w(Y^{\hat{i}}, Y^{\hat{j}})]$$
$$= 0\,.$$

Thus,

$$\mathbb{P}\left(|\hat{q}(x) - \mathbb{E}[\hat{q}(x)]| > \epsilon\right) \leqslant \frac{2}{k\epsilon^2}\,,$$

which converges to 0 for any fixed $\varepsilon > 0$, as long as $k \to \infty$. This concludes the proof.

D Proof of Proposition 1

Proposition 2. *Let $\psi_\pi(\mathcal{P}) = V(\pi)$ be our estimand.*

1. *For $\pi(x) = 1$ the constant policy, $\psi_\pi = \psi_W$ is the win proportion, i.e. the likelihood that the primary intervention is preferable to the alternative. Its EIF writes as; for $o = (X, T, Y)$:*

$$\begin{aligned}\phi_W(\mathcal{P})(o) = &-\psi_W(\mathcal{P}) + \mathbb{E}\left[w(Y(1,X_i)|Y(0,X_i))|X_i = X\right] \\ &+ \frac{T}{e(X)}\left(\mathbb{E}\left[w(Y|Y(0,X_i))\right] - \mathbb{E}\left[w(Y(1,X_i)|Y(0,X_i))|X_i = X\right]\right) \\ &+ \frac{1-T}{1-e(X)}\left(\mathbb{E}\left[w(Y(1,X_i)|Y)\right] - \mathbb{E}\left[w(Y(1,X_i)|Y(0,X_i))|X_i = X\right]\right)\,,\end{aligned}$$

   where $e(x) \overset{def}{=} \mathbb{P}\left(T_i = 1|X_i = x\right)$.

2. *For $\pi(x) = 0$ the constant policy, $\psi_\pi = \psi_L$ is the loss proportion, i.e. the likelihood that the primary intervention is detrimental relative to the alternative, and its EIF writes as; for $o = (X, T, Y)$:*

$$\begin{aligned}\phi_L(\mathcal{P})(o) = &-\psi_L(\mathcal{P}) + \mathbb{E}\left[w(Y(0,X_i)|Y(1,X_i))|X_i = X\right] \\ &+ \frac{T}{e(X)}\left(\mathbb{E}\left[w(Y(0,X_i)|Y)\right] - \mathbb{E}\left[w(Y(0,X_i)|Y(1,X_i))|X_i = X\right]\right) \\ &+ \frac{1-T}{1-e(X)}\left(\mathbb{E}\left[w(Y|Y(1,X_i))\right] - \mathbb{E}\left[w(Y(0,X_i)|Y(1,X_i))|X_i = X\right]\right)\,.\end{aligned}$$

3. *For general policy π, the EIF of ψ_π is:*

$$\phi_\pi(\mathcal{P})(o) = -\psi_\pi(\mathcal{P}) + \pi(x)(\phi_W(\mathcal{P})(o) + \psi_W(\mathcal{P})) + (1 - \pi(x))(\phi_L(\mathcal{P})(o) + \psi_L(\mathcal{P}))$$

Proof of Proposition 2. We have:

$$\begin{aligned}\Psi_W(\mathcal{P}) &\overset{\text{def}}{=} \mathbb{E}\left[\mathbb{E}\left[w(Y_i(1,X_i)|Y_i)|X_i, T_i = 0\right]\right] \\ &= \int \mathrm{d}\mathbb{P}_{X_i}(x)\mathrm{d}\mathbb{P}_{Y_i|X_i=x,T_i=0}(y)\mathrm{d}\mathbb{P}_{Y_i|X_i=x,T_i=1}(y')w(y'|y) \\ &= \int p(x)p(y|x,0)p(y'|x,1)w(y'|y)\mathrm{d}x\mathrm{d}y\mathrm{d}y' \\ &= \int p(x)\frac{p(x,0,y)}{p(x,0)}\frac{p(x,1,y')}{p(x,1)}w(y'|y)\mathrm{d}x\mathrm{d}y\mathrm{d}y'\,,\end{aligned}$$

where p is the density of $\mathcal{P}$ with respect to the Lebesgue measure if $\frac{\mathrm{d}\mathcal{P}}{\mathrm{d}\lambda} < \infty$, and this expression naturally extends in the general case. Let $h \in (0,1)$ (that will be taken infinitesimally small) and $o = (X, T, Y)$. Let $\mathcal{P}_h = (1-h)\mathcal{P}_h\delta_{X,T,Y}$ be the mixture of these two distributions. We have:

$$\Psi_W(\mathcal{P}_h) = \int p_h(x)\frac{p_h(x,0,y)}{p_h(x,0)}\frac{p_h(x,1,y')}{p_h(x,1)}w(y'|y)\mathrm{d}x\mathrm{d}y\mathrm{d}y'\,,$$

where p_h is the density of $\mathcal{P}_h$. Note that we have $p_h = (1-h)p + h\delta_{X,T,Y}$, leading to:

$$\begin{aligned}\Psi_W(\mathcal{P}_h) = (1-h)\Psi_W(\mathcal{P}) + o(h) + h\int p(y|X,0)p(y'|X,1)w(y'|y)\mathrm{d}y\mathrm{d}y' \\ + h\frac{1-T}{1-\pi(X)}\int p(y'|X,1)w(y'|Y)\mathrm{d}y' - h\frac{1-T}{1-\pi(X)}\int p(y|X,0)p(y'|X,1)w(y'|y)\mathrm{d}y\mathrm{d}y' \\ + h\frac{T}{\pi(X)}\int p(y|X,0)w(Y|y)\mathrm{d}y - h\frac{T}{\pi(X)}\int p(y|X,0)p(y'|X,1)w(y'|y)\mathrm{d}y\mathrm{d}y'\,.\end{aligned}$$

Because $\Psi_W(\mathcal{P}_h) = \Psi_W(\mathcal{P}) + h\Phi(\mathcal{P})(o) + o(h)$, we can identify Φ_W as in Proposition 1. The proof for Φ_L and general policies Φ_π proceeds exactly in the same way. □

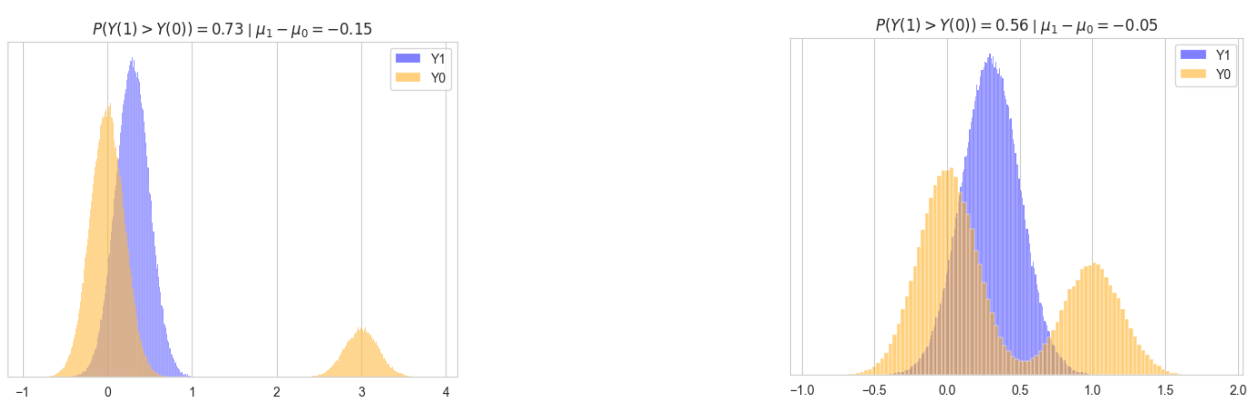
E DETAILED EXPERIMENTAL SECTION

E.1 Synthetic simulation details

We explore in our synthetic experiments the benefits of applying our CPTE methodology to preference-based policy learning and evaluation. We take as a use case PNS over a continuous outcome, $\mathbb{E}\left[w(Y_i(1), Y_j(0))|X_i = X_j = x\right] = \mathbb{P}\left(Y_i(1) > Y_j(0)|X_i = X_j = x\right)$.

In some cases a CATE-based policy suffices to optimize a CPTE target. For example, CATE estimation is sufficient to learn a binary PNS preference because the Bernoulli distribution is characterized by its expectation. However, complex potential outcome distributions can result in divergence between an optimal CATE policy and an optimal preference-based policies, such as PNS. We give a real-world example below.

To create our experimental data, we simulate a mix of binary and continuous features. The potential outcomes are generated from a combination of linear effect modifiers and a complex treatment effect. We present two scenarios; one with homogeneous treatment effects and the other with a simple heterogeneous treatment effect. The simulation maintains Assumptions 1 to 3. To test our CPTE methodology against classical CATE-based policies, the complex treatment effect was simulated such that the two policies disagree on every unit. This can occur when one of the potential outcome distributions has a heavy tail. Real-world applications of this could be risky surgeries that benefit certain individuals greatly but are harmful for most patients. In this case, the minority of patients who benefit greatly may skew the average, making the riskier surgery appear more beneficial than a more conservative option whose likelihood of benefiting the patient is better. A CATE policy would recommend the risky surgery; a CPTE policy would not.
Our simulation mimics this scenario by drawing one potential outcome from a gaussian distribution and the other potential outcome from a bimodal normal distribution. Figures 2a and 2b illustrate two examples of potential outcome distributions for which there is divergence between a PNS-based policy and a CATE-based policy. Other more subtle examples can be simulated by using heavy-tail distributions such as Gamma and Pareto.



(a) Large divergence

(b) Small divergence

Figure 2: Potential outcome distributions

In the homogeneous setting, all units potential outcomes are drawn from the same potential outcome distribution modified by a linear combination of the features. In the heterogeneous setting, an observed binary feature was used to decide which potential outcome was assigned the bimodal distribution. We include more details in appendix E.1.1.

For each simulation we perform policy learning and evaluation with three distributional learning paradigms - KNN, linear regression and random forest. These are evaluated on a held-out set (N=10000) generated using a random seed that is not used for any of the training data simulations. Training data is simulated 50 times for each N. We sample a training set for each $N \in [30, 100, 1000, 10000]$ and train our distributional and baseline estimation methods. Using their estimates for the conditional PNS, we evaluate three policy rules. The first, OTR (section 5.2), uses the $\delta(x) > 0$ rule; $\delta(x) = q_W(x) - q_L(x)$ for CPTE and $\delta(x) = \mathbb{E}[Y_i(1) - Y_i(0)|x]$ for CATE. The second policy rule, policy value maximization (section 5.3) uses a logistic regression model to learn $f : X \to \mathbf{I}(\delta(x) > 0)$ over the training data. The policy follows the prediction. The third policy rule also uses a logistic regression model to decide policy, but learns over the EIF correction of $\delta(x)$ (eq. (13)).

We compare our CPTE approach using distributional estimation to a standard approach using CATE, which serves as our baseline. Distributional KNN (D-KNN), estimating $\mathcal{P}_{Y_i|T_i,X_i}$ is compared to regular KNN, estimating $\mathbb{E}[Y_i|T_i, X_i]$. Likewise, Distributional Linear Regression (DLR) using quantile regression is compared to regular ridge regression; Distributional Random Forest (DRF) is compared to regular random forest.

E.1.1 Synthetic Data Generating Process

We generate features $X \in \mathbb{R}^{10}$ using 8 continuous variables and two binary variables. We randomly sample linear coefficients to create a shared baseline dependent on X for the potential outcomes. In the observational setting, we choose three features (one discrete and two continuous) to act as confounders and use a linear transformation to create the logit for treatment arm selection. We ensure that overlap exists and set the expected treatment rate to 0.5.

- Continuous features: $X_{cont} \sim \mathcal{N}(0, 1)$
- Discrete features: $X_{disc} \sim Ber(0.5)$
- Homogeneous treatment assignment: $T \sim Ber(0.5)$
- Heterogeneous treatment coefficients: $\beta_T \sim U(0.2, 0.6)$ for one discrete and two continuous features acting as confounders, otherwise 0.
- Heterogeneous treatment assignment: $T \sim Ber(X^T \beta_T)$
- Baseline outcome coefficients: $\beta_Y \sim U(0.1, 0.5)$ for all features.
- Shared baseline Y: $Y_{baseline} = X^T \beta_Y$

We define two distributions for the potential outcome space: a bimodal mixture of Gaussians and a simple Gaussian distribution. The bimodal distribution has a higher expected value than the simple Gaussian, as visualized in fig. 2a and fig. 2b. However, the probability that a random variable of the simple Gaussian distribution is higher than the bimodal distribution is above 0.5. We let the $b \in [0, 1]$ parameter control how many are part of the lower Gaussian and $1 - b$ how many are part of the higher Gaussian. Referencing the risky surgery example mentioned above, $1 - b$ represents how many patients are in the minority group that greatly benefited from the risky surgery.
Potential outcome parameters:

- Shared variance: $\sigma = 0.2$
- Simple Gaussian parameters: $\mu_s = 0.3$
- Simple Gaussian: $y_{simple} \sim \mathcal{N}(\mu_s, \sigma^2)$
- Mixture of Gaussians parameters: $\mu_{m,low} = 0, \mu_{m,high} = 3, b = 0.85$

- Mixture of Gaussians: $y_{mixture} = b * \mathcal{N}(\mu_{m,low}, \sigma^2) + (1-b)\mathcal{N}(\mu_{m,high}, \sigma^2)$

We create the potential outcomes by combining the linear baseline outcome with these distributions. In the homogeneous setting,

$$Y(1) = y_{baseline} + y_{simple}$$

$$Y(0) = y_{baseline} + y_{mixture}$$

Therefore, for all units the optimal PNS policy chooses T=1 and the optimal CATE policy chooses T=0.

We denote one of the binary features as X_0. In the heterogeneous setting,

$$Y(1) = y_{baseline} + \mathbf{I}(X_0 = 1)y_{simple} + \mathbf{I}(X_0 = 0)y_{mixture}$$

$$Y(0) = y_{baseline} + \mathbf{I}(X_0 = 0)y_{simple} + \mathbf{I}(X_0 = 1)y_{mixture}$$

Therefore, the optimal PNS and CATE policies depend on the binary feature X_0.

E.1.2 Experiment Details

Evaluation. Evaluation was performed on a large hold-out set (N=50,000). X and $y_{simple}, y_{mixture}$ were generated using a random seed not used by any of the training sets. The parameters for β_y were generated using a random seed shared across all datasets. Therefore, $y_{baseline}$ comes from the same distribution for both the evaluation data and the training data.
Training. For each $N \in \{30, 100, 1000, 10000\}$ the training data was simulated 50 times. This allows us to observe the performance of the models over exponentially growing datasets. Each time X and $y_{simple}, y_{mixture}$ were generated using a different random seed.
For each learning paradigm - KNN, linear regression and random forests - we train a distributional model and a baseline model. The distributional model estimates the distribution of the potential outcomes, $\hat{Y}_i(1), \hat{Y}_i(0)|X_i = x$ under Assumptions 1 to 3. Distributional KNN (DKNN) uses the nearest neighbors to approximate the distribution of the potential outcomes. We test nearest neighbor (k=1) and logarithmically growing k's. The distributional linear and forest regressors approximate the inverse cdf using quantiles and then sample 1000 times from the approximate potential outcome distribution. We denote the number of samples used as S. The baseline learner estimates the expected value of the potential outcome under Assumptions 1 to 3. These estimations allow us to perform policy learning and evaluation.
Policies are learned in three ways:

- Plug-in of the OTR (section 5.2)
- Plug-in Policy Value Optimization (section 5.3)
- EIF Policy Value Optimization (see section 6)

Policy Learning & Policy evaluation For each learning paradigm, we compare the CPTE-policies to the CATE-policies using the oracle policy value on the evaluation set. This allows us to understand which policy perform best. The full results can be found in fig. 3 and are analyzed in the next section.

Policy evaluation was performed by using each paradigm's estimates on the evaluation set and comparing the optimal policy's value using them. The CATE policy value was calculated by plugging in the estimated expected values into the preference function (definition 1).

Model Parameters The baseline linear regression model is Ridge with alpha=1.0.

The random forest parameters depend on the size of N. These were found by performing a hyperparameter search for each N using a random seed not used during training. This allowed us to understand reasonable parameters given different training set sizes.

Table 1: Random Forest Regressor hyperparameter settings by sample size N.

Condition on N	Max Depth	Max Features	Min Samples Split	# Trees ($n_{\text{estimators}}$)
$N \geqslant 10000$	25	0.35	5	500
$100 < N < 10000$	15	0.50	7	400
$N \leqslant 100$	15	0.60	5	50

E.1.3 Synthetic Experiments: Detailed Results

Policy learning. As we see in Figure 3, the distributional approaches tend to converge to the optimal policy value. We also note that the variance, represented by bootstrapped 95% confidence intervals, tends to get smaller for the distributional approaches. In contrast, the baseline learners tend to converge to the incorrect CATE policy and the variance tends to be larger. The simulation is linear; resulting in the quick convergence of all Distributed Linear Regression (DLR) policies to the optimal policy in all settings and for all policy learning methods. The baseline plug-in and direct policy value optimization converge to the optimal CATE policy, the opposite of the optimal CPTE policy. We note that the EIF PNS correction of the CATE facilitates above-random performance in some of the settings. The Distributed KNN (DKNN) policies also converge towards the optimal policy value. Despite being misspecified to the linear simulation, the DKNN plug-in policy correctly estimates the PNS, allowing the correct policy to be learned as a logarithmic rate as anticipated by Theorem 2. Unlike DKNN, the Distributional Random Forest (DRF) method struggles to overcome its misspecification in the heterogeneous setting. Although not converging to the CATE policy like the baseline random forest, it struggles to learn the correct distribution. This results in a plug-in that stays close to a random policy because the heterogeneity is equally balanced. Despite this inadequacy, the policy value optimization methods converge to the optimal policy.

Policy evaluation. Beyond the distributional method's ability to learn policies over complex potential outcome distributions, they also enable correctly-specified policy evaluation for complex preference functions like PNS. We evaluate the optimal policy in term of PNS using PNS estimates from the distributional and baseline models. As can be seen in fig. 4, the distributional estimates converge towards the optimal policy. The correctly specified Distributional Linear Regression converges quickly to the correct policy value. DKNN and DRF tend towards the correct policy value. In contrast, the baseline estimation using the expected potential outcomes diverge to 0 and 1. We note that the true policy value is 0.72.

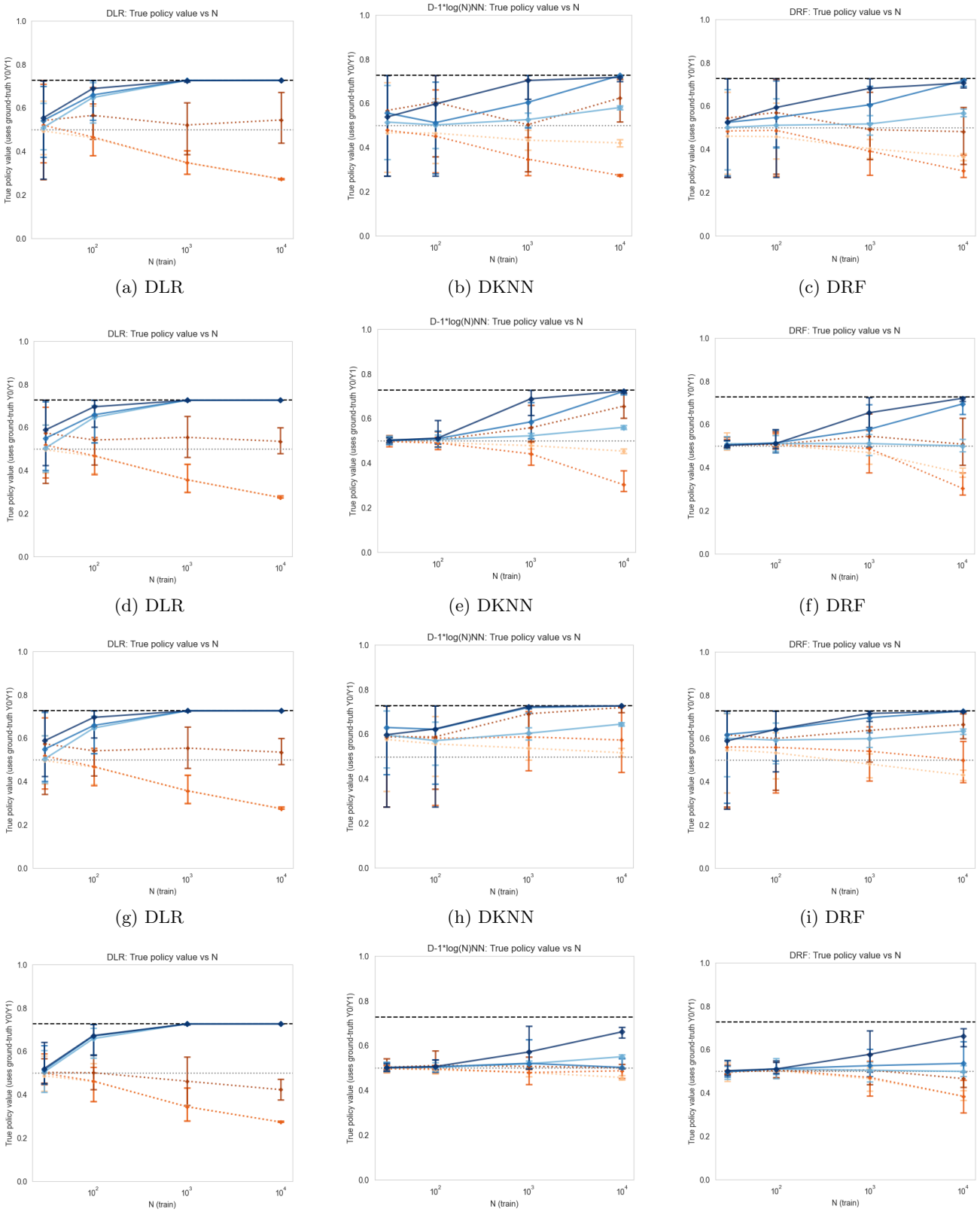
Correlation between potential outcomes. We also examine when the assumption made in Theorem 1, i.e. all structural treatment effect modifiers must be observed, is violated. This assumption is unlikely to hold; these experiments can help us understand the impact of being far from this assumption. We test this by introducing correlations between the potential outcomes that are not explained by the available features. Continuing the surgery example, patients for whom the risky surgery is successful may have a genetic component that also affects their reaction to the more conservative treatment.

We show that in our simple synthetic setting, policy learning appears to be more robust to the violation of this assumption than policy evaluation (Figure 5). For the policy learning, we focus on the successful DKNN and distributional linear regression in the difficult heterogeneous setting. We add significant amount of positive and negative correlation between the potential outcomes with empirical Pearson correlations of 0.7 and -0.7 respectively.

Policy learning is hardly affected. The visual in Figure 6 shows how the correlation only affects a small part of the population (blue concentric circles) which cross the decision boundary (red line.) However, the policy value estimate is significantly effected. We show in Figure 7 policy evaluation using the correctly specified linear models. Under correlation, the policy values converge to the wrong policy value. They converge to the policy value without correlation.

E.2 Semi-Synthetic Simulation details

Validating causal-inference methods is difficult because the counterfactual ground-truth necessary for validation is never observed. This phenomena is called the *fundamental challenge of causal inference* (Holland, 1986). The natural solution is to verify methods under ideal circumstances, such as random control trials. Random control



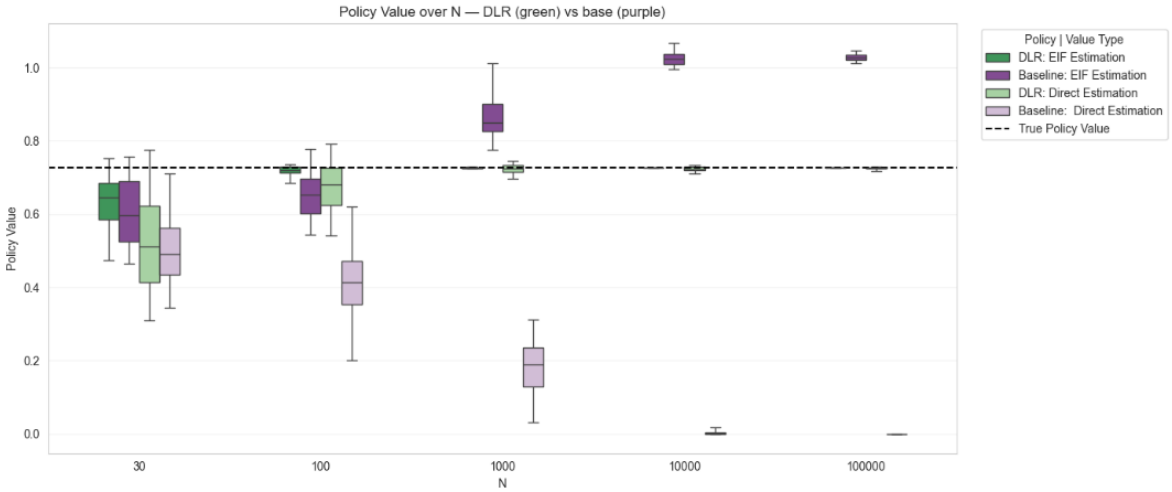
(a) DLR (b) DKNN (c) DRF

(d) DLR (e) DKNN (f) DRF

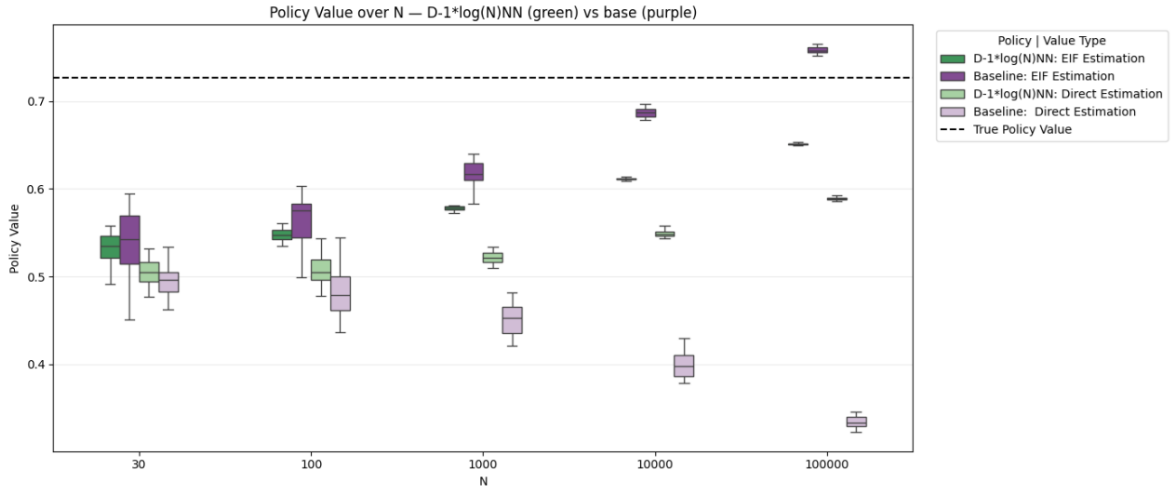
(g) DLR (h) DKNN (i) DRF

(j) DLR (k) DKNN (l) DRF

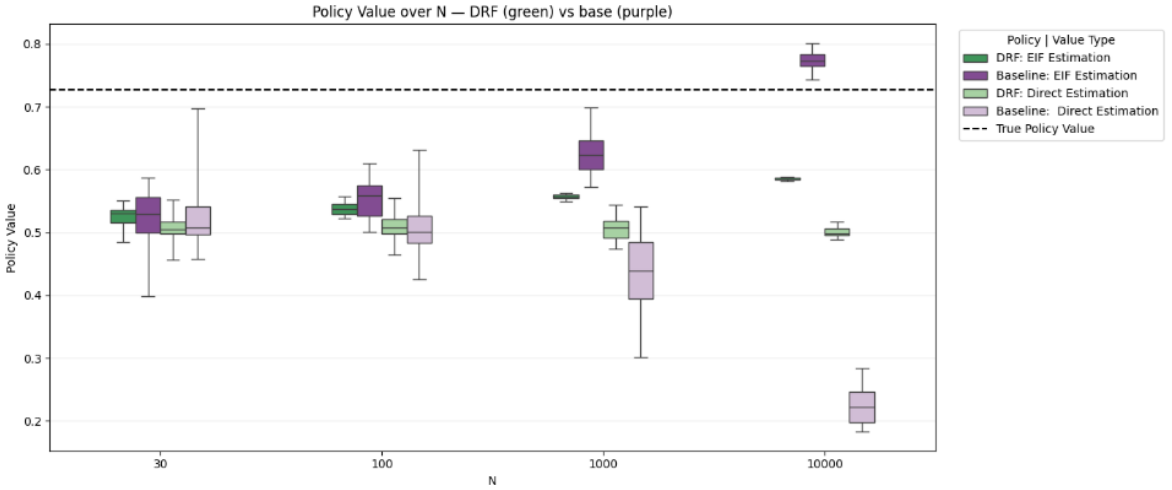
Figure 3: Policy Learning with PNS policy value, no correlations. (a–c) Homogeneous RCT setting; (d–f) Heterogeneous RCT setting; (g-i) Homogeneous Observational setting; (j-l)Heterogeneous Observational setting. **Legends** as in Figure 1a.



(a) DLR

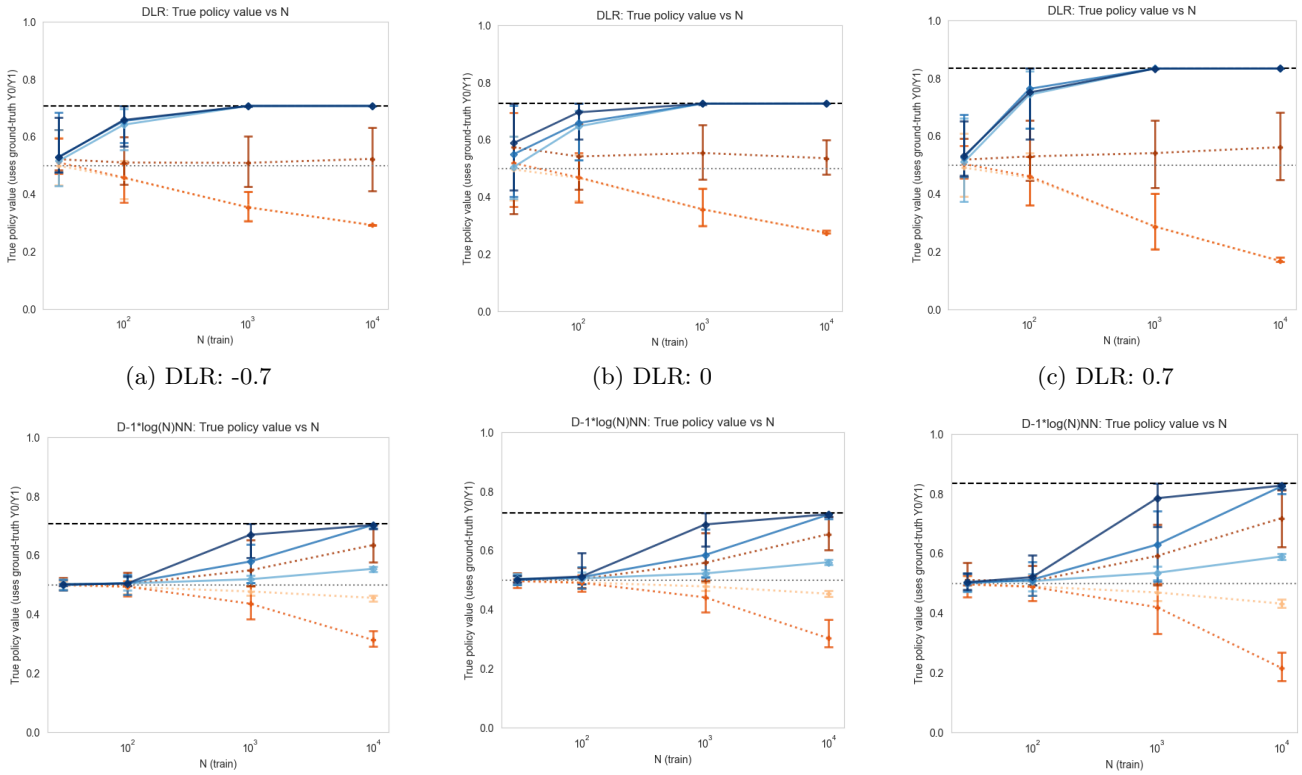


(b) DKNN



(c) DRF

Figure 4: Policy Evaluation of optimal policy in the synthetic, heterogeneous RCT setting



(a) DLR: -0.7 (b) DLR: 0 (c) DLR: 0.7

(d) DKNN: -0.7 (e) DKNN: 0 (f) DKNN: 0.7

Figure 5: Policy Learning with correlation between potential outcomes in the synthetic, heterogeneous RCT setting. (a–c) Linear Regression; (d–f) DKNN. **Legends** as in Figure 1a.

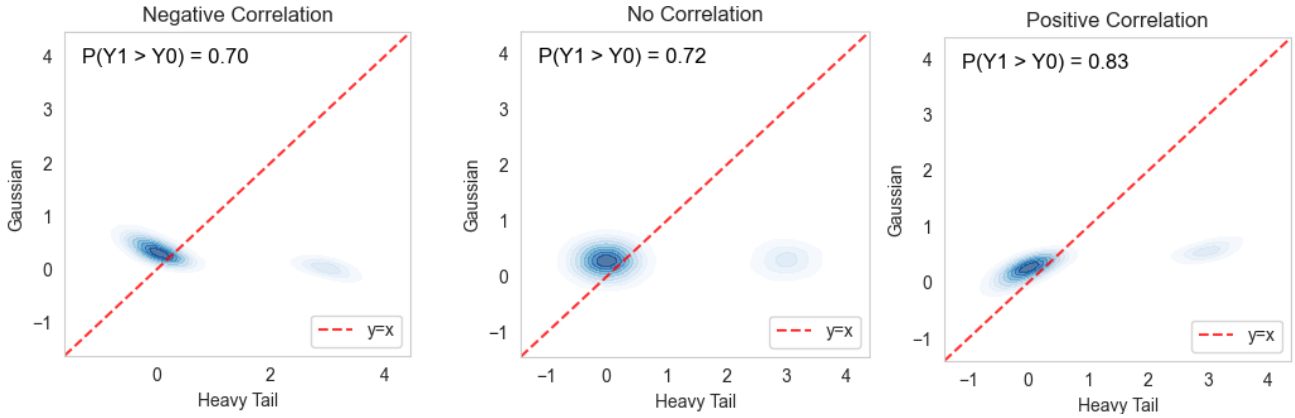


Figure 6: Visualizing PNS, $P(Y1 > Y0)$, with correlated potential outcomes $Y1$ and $Y0$. In our experiment, $Y1$ is Gaussian (y-axis) and $Y0$ has a heavy tail (x-axis).

trials provide unbiased estimates for CATE-estimates because they are unconfounded by construction. While CPTE-estimates also benefit from an RCT's inherent identifiability, they are still statistically problematic. Means can be reasonably estimated for small substratum. However, estimating the conditional distribution necessary for CPTE requires more data. This is especially true because the CPTE is particularly suited to tasks with complex conditional potential outcome distributions, as we have shown for PNS in the simulation above.

Therefore, to further validate our methods we turn to a semi-synthetic experiment. We use the Tennessee STAR study (Word et al., 1990) for the starting point. The Tennessee Student/Teacher Achievement Ratio (STAR) study is a randomized experiment intended to measure the effect of class size on student test scores, namely math, reading and listening. Launched in 1985, the study followed students in grades K-3 recording their grades at the end of each academic year. Students were randomly assigned to one of three classroom types - small (13-17 students), regular (22-25) and regular with a full-time aide (22-25). Each school housed at least one classroom of each type. (Achilles et al., 2008). We apply our methods to the 'win' function, a preference function (definition 1), which is often aggregated into various 'win' statistics (Dong et al., 2023). The 'win' function captures preference over hierarchical outcomes.
We work with a primary, binary outcome and a secondary, continuous outcome. This scenario is common in medical literature, where the primary outcome is often survival - binary or time-to-event - as the primary outcome and a secondary, continuous outcome such as blood pressure. Our primary outcome is experiment attrition. We looked at kindergarten students found in the study and defined attrition as those students who are not found in the study in the first grade. Our secondary outcomes are the student's math scores. We define a 'win' as students who are found in the study in first grade and have a high math score.
After removing students without a math score, we there are 1,762 student in small classrooms, 2,032 in regular classrooms and 2,077 in regular classrooms with full time aides. 454 students were removed due to a missing math score. We combine the two treatment arms with regular classrooms because the 'win' function is only defined between two treatments. Although this may violate SUTVA, one of the study's findings was that the full-time aides were not significant (Achilles et al., 2008; Krueger, 1999).

As stated previously, calculating the conditional 'win' function empirically over the RCT is statistically problematic, as many small substratum result in poorly estimated 'win' functions despite being causally unbiased. Therefore, we create synthetic outcomes. We strive to replicate the original outcome distributions. We do so by training an MLP over the outcomes and then sampling from the MLP using MC-dropout (Gal and Ghahramani, 2015), resulting in a Bayesian approximation of conditional distributions.

We then use these estimates of the conditional distributions as the ground-truth for the optimal policy and the optimal policy value. We evaluate our methods in a similar fashion to the synthetic experiments, training distributional and baseline models. We then use these models' predictions as plug-in estimators for the 'win' function and predict the policy using the same policy learning methods as in the synthetic experiment.
In the synthetic experiments, we showed the advantages of a PNS-based policy over a mean-based policy for complex potential outcome distribution. The advantage of the 'win' function is evident even without that complexity. Because the 'win' function is composed of hierarchical outcomes, a successful 'win'-based policy discerns between units for which the benefits of the secondary outcome outweigh the risks in the primary outcome. The baseline learners, which estimate the expected values of outcomes, will likely only prioritize the primary outcome. In the binary case, only when the estimated expectations are equal will the secondary outcome be considered.

We present our results on the semi-synthetic STAR data using the original features and our simulated outcomes. In contrast to the synthetic experiment, this more realistic simulation contains preferences close to the decision boundary, making policy learning more difficult.

We find in fig. 8 that as the size of our training data grows, the distributional methods best learn the optimal policy. In our graphs, our results with the oracle policy value (black) and by a random policy (gray). We also add a benchmark for the optimal primary outcome policy (brown), i.e. the policy value if we would base our policy only on the primary outcome, not the 'win' preference function. We find that all of the baseline plug-in policies (yellow) do not even achieve the optimal primary outcome policy. The distributional policies (blues) and the direct policy value optimization policy for the baseline overtake the optimal primary outcome policy. These results indicate the power of distributional estimators to optimize CPTE functions. In contrast, the plug-in baseline policy barely beats an 'preference' value of 0.5. This indicates the failure of CATE estimates performance on CPTE outcomes like 'win' statistics.

We also experiment with correlated potential outcomes in the semi-synthetic setting. Our experiments show that our models and the baseline models are not sensitive to this violation of the assumptions for Theorem 1. See fig. 9.

E.2.1 Data Collection and Processing

Creating the cohorts We use the STAR dataset as the starting point for our semi-synthetic experiment as it appears in the AER R package (Kleiber and Zeileis, 2008). We select students in kindergarten and create our binary outcome by checking who appears in the 1st grade cohort. Data is missing from the *math* outcome (454 students). The missing *math* outcome was dropped. We split the students into two intervention arms, small classrooms and regular classrooms. The latter arm includes two groups, regular classrooms without aides and those with full-time aides. Although, this may be a violation of SUTVA, the study's findings indicate that the full-time aides were not significant (Achilles et al., 2008; Krueger, 1999). This resulted in two groups, the small classroom group (1,762 students) and the regular classroom group (4,109 students).

Features Relevant features used to create the simulated outcomes and for training are: student gender, student ethnicity, teacher gender, teacher ethnicity, years of teaching experience, teacher's highest degree, where the teacher is found in their career (e.g. apprentice, probation, level 1, etc.), school location (rural/inner-city/urban/suburban) and whether the school offered free/reduced-fee lunches. The last variable is an indicator of low-income family background (Achilles et al., 2008). All the variables are categorical with the exception of the years of teaching experience. Categorical variables were one-hot encoded and missing values were assigned zero to all categories. Missing values in years of experience were assigned the median.

Table 2: Missing Variables

Variable	Variable Type	Missing Values	Pre-processing
Math	outcome	454	Removed
Student Ethnicity	feature	1	Assigned zero in all categories
Free/Reduced-Fee Lunch	feature	17	Assigned zero in all categories
Degree	feature	21	Assigned zero in all categories
Career Ladder	feature	541	Assigned zero in all categories
Teacher Ethnicity	feature	60	Assigned zero in all categories
Years of Teaching Experience	feature	21	Median

Outcome Simulation An MLP was trained for each intervention arm and outcome. We summarize the hyper-parameters and evaluation metrics of each model. Dropout of 10% was used for all models in training

Table 3: Summary of deep model models

Model	Hidden Dimensions	Epochs	Batch Size	Learning Rate	Brier Score	RMSE (std)
Small – Dropout	[64, 64]	35	128	5×10^{-4}	0.17	–
Regular – Dropout	[64, 64]	34	128	5×10^{-4}	0.19	–
Small – Math Score	[64, 64, 64]	100	32	1×10^{-4}	–	48.5 (49.5)
Regular – Math Score	[64, 64, 64]	100	32	1×10^{-4}	–	45.1 (46.7)

After training the models, we sample using MC-dropout (Gal and Ghahramani, 2015) with a drop-out rate of 5%. For each unit, we sample 1000 potential outcomes for each intervention arm. These simulations induce a difficult, non-trivial policy learning problem. Most units' simulated 'win' preference function are between 0.4 and 0.6. We compare the optimal policies for the primary and secondary outcome:

Table 4: Comparing Optimal Policies

Policy	Treat w/ smaller	Agreement with Win Policy	Win Policy Value
Win (Dropout, Math) Policy	77%	100%	0.59
Dropout Policy	67%	70%	0.54
Math Score Policy	75%	92%	0.58

As can be seen from table 4, the policy prioritizing the secondary outcome is close to the optimal policy. This is due to the lack of significant treatment effect on the binary outcome. However, this poses a difficulty to learning the optimal policy in the 'win' setting. The hierarchical ordering of 'winning' forces the policy to first take into consideration the primary outcome.

Training & Evaluation The training and evaluation are similar to the synthetic simulation. The primary difference is the inclusion of a binary, primary outcome. Binary outcomes are completely parameterized by their Bernoulli parameter, which is the expected value. There is no need for our distributional methods on the primary outcome. Therefore, we perform model selection on the training data and use the same model for all distributional and baseline methods. We found a KNN classifier with 11 neighbors to offer the best brier score (0.17). Our sample size is also limited by the size of the kindergarten cohort. We divide our data into a hold-out set containing 1,500 students (25%) and our pool of training data is the remaining 4,403 students. We randomly sample training data with $N \in [30, 100, 1000, 4403]$ and train the estimation models with their derivative policy learning models. We repeat this process fifty times for each model and N.

E.2.2 Correlation

We also explore the impact of correlation between potential outcomes in our semi-synthetic study. To do so, we employ the Iman-Conover method (Iman and Conover, 1982) commonly used to induce correlation between variables in simulations (Huang et al., 2012; Tan et al., 2025; Faries et al., 2020). This method is distribution free and preserves the marginal distributions. Instead of changing the values, it reorders them in such a way that induces an approximation of the desired correlation. We use this method to induce correlation between the potential outcomes of each outcome. Therefore, the marginal potential outcome distributions are identical between the three experiments. For each experiment, we take a single sample from the potential outcomes of each unit and assign the observed outcome based on the unit's treatment arm.

We evaluate our models when there is positive or negative correlation between the potential outcomes, results are available in Table 5. For the positive correlation, the desired correlation was set at 0.5 for each outcome. The induced correlation was 0.28 for the binary potential outcomes and 0.47 for the continuous potential outcomes. For the negative correlation, the desired correlation was set at -0.5 for each outcome. The induced correlation was -0.22 for the binary potential outcomes and -0.41 for the continual potential outcomes.

Table 5: Comparing Optimal Policies

Correlation (Induced)	Policy	Treat w/ smaller	Agreement w/ Win Policy	Win Policy Value
-0.5 (-0.22/-0.41)	Win Policy	73%	100%	0.59
	Dropout Policy	67%	62%	0.53
	Math Score Policy	75%	98%	0.59
0	Win Policy	77%	100%	0.59
	Dropout Policy	67%	70%	0.54
	Math Score Policy	75%	92%	0.58
0.5 (0.28/0.47)	Win Policy	74%	100%	0.65
	Dropout Policy	67%	63%	0.55
	Math Score Policy	75%	99%	0.65

Similarly to the synthetic experiment, the correlations change the optimal policy slightly (8%) and the positive correlation raises the value of the optimal policy value. We show in fig. 9 that our models are only slightly affected. Because the positive correlation causes the units 'win' functions to be more extreme, we notice that as

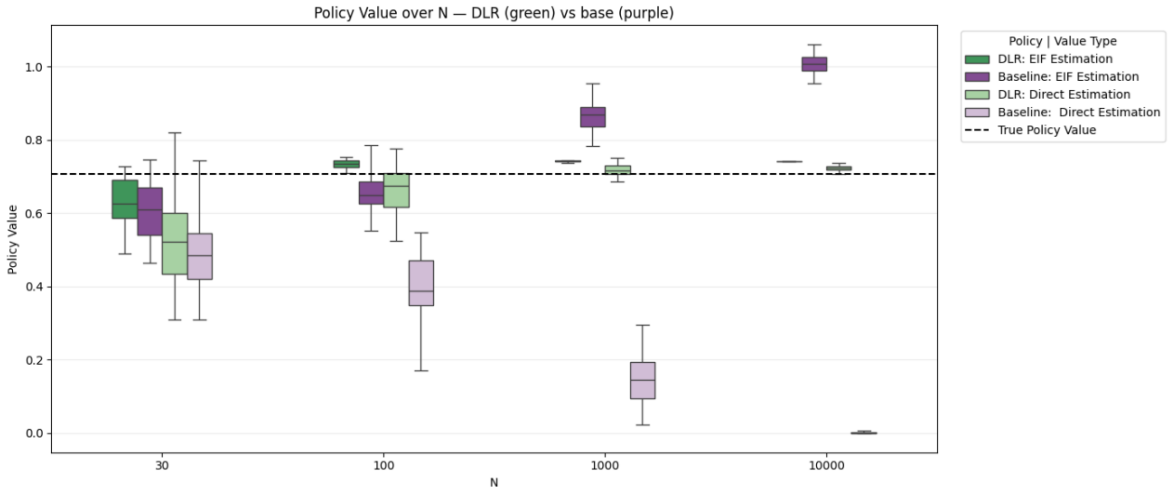
the models better learn the correct policy they gain significant policy value.

E.2.3 Impact of the number of neighbors in KNN

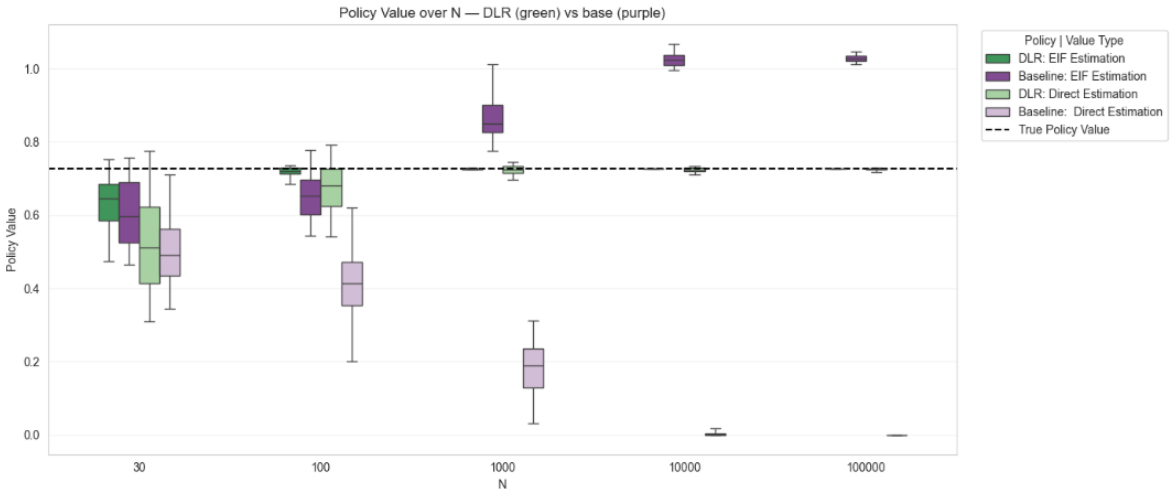
Figure 10 shows the impact of adding more neighbors for policy learning. As predicted by our theory, the plug-in OTR (Section 5.2) requires $k \to \infty$ to converge. Indeed, for $k = 1$ the plug-in method performs as poorly as the baseline plug-in (almost as poorly as random guesses), while increasing k leads to higher policy values, and faster with higher k. However, and quite interestingly, this is not the case for policy value optimization methods (Section 5.3 and appendix A.3) due to the fact that values are averaged over the whole population. In fact, adding more neighbors does not seem to impact much convergence.

E.2.4 Policy Agreement - Optimizing 'Win' vs. Primary outcome vs. Secondary outcome

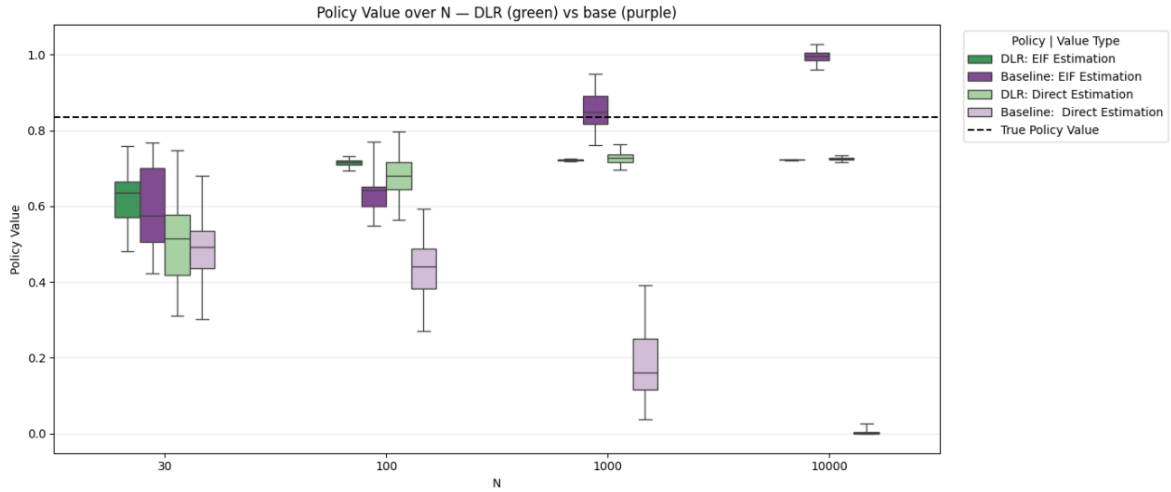
Lastly, we compare the *policy agreement i.e.*, the percentage of agreement between our learned policies and a given oracle policy that is either *(i)* the optimal policy learned with oracle information using the hierarchical outcomes, *(ii)* the optimal policy learned with oracle information that only optimizes the primary outcome, and *(iii)* the optimal policy learned with oracle information that only optimizes the secondary outcome. Results are reported in Figure 11. We show results for the DLR method. These results are also consistent for KNN and DRF.



(a) DLR - negative correlation

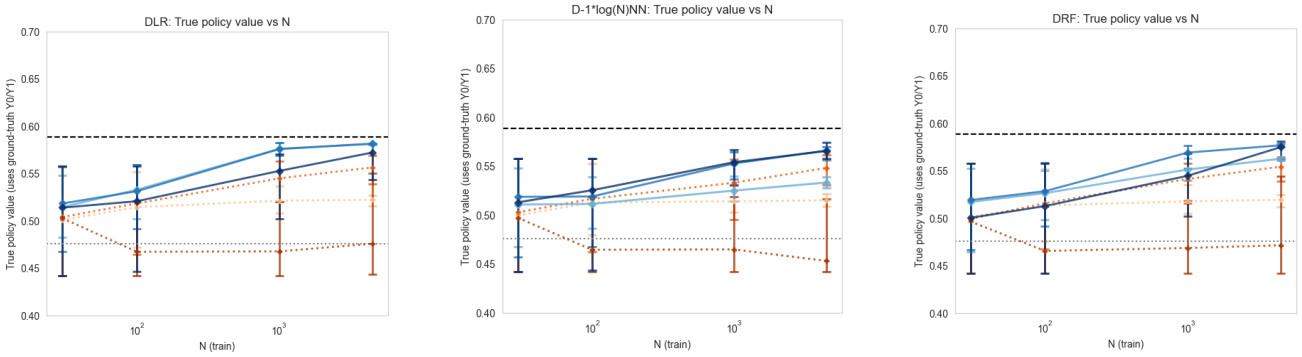


(b) DLR



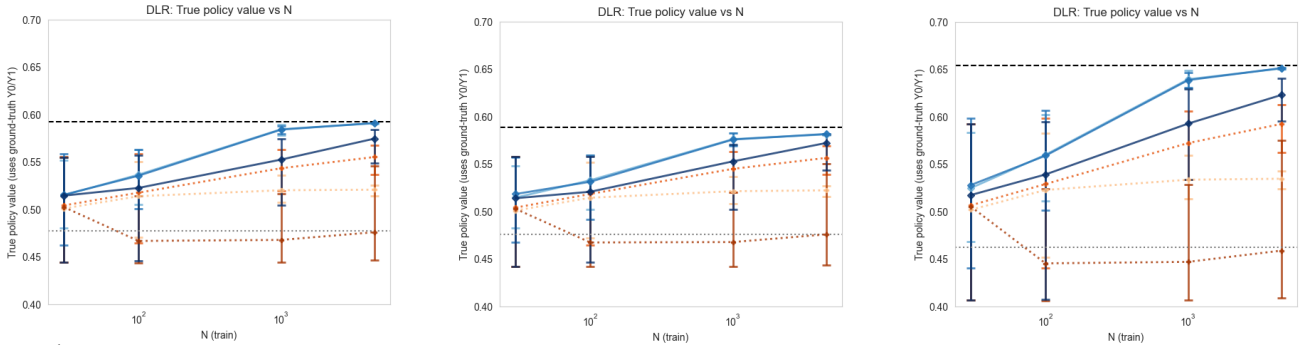
(c) DLR - positive correlation

Figure 7: Policy Evaluation of optimal policy with correlation in the semi-synthetic setting

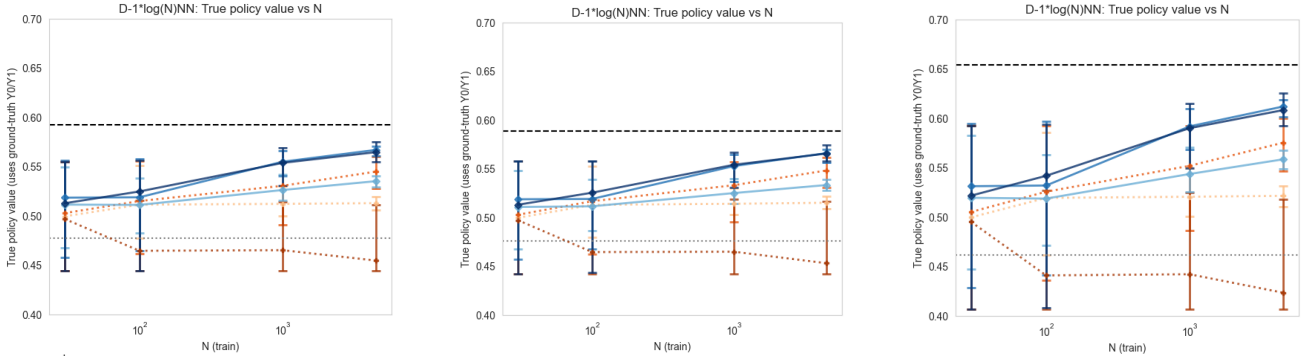


(a) DLR (b) DKNN (c) DRF

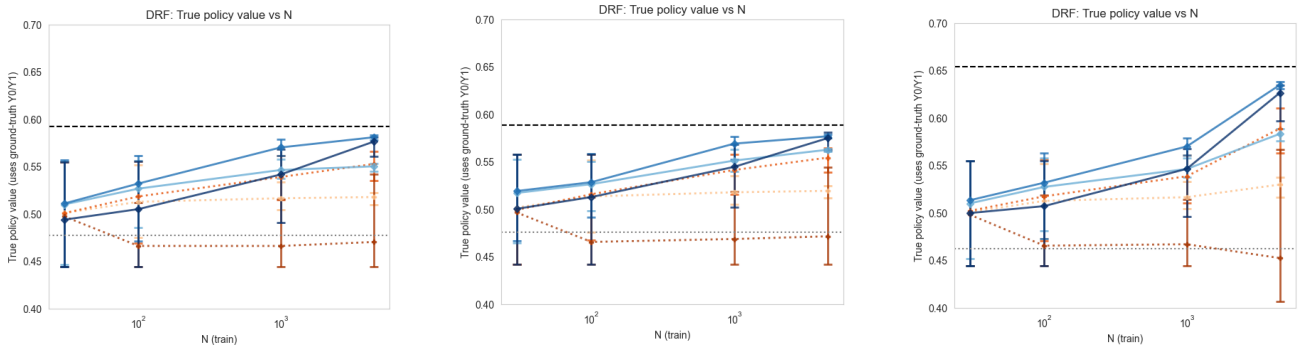
Figure 8: Policy Learning with win policy value in the semi-synthetic setting. **Legends** as in Figure 1a.



(a) DLR: -0.5 (b) DLR: 0 (c) DLR: 0.5

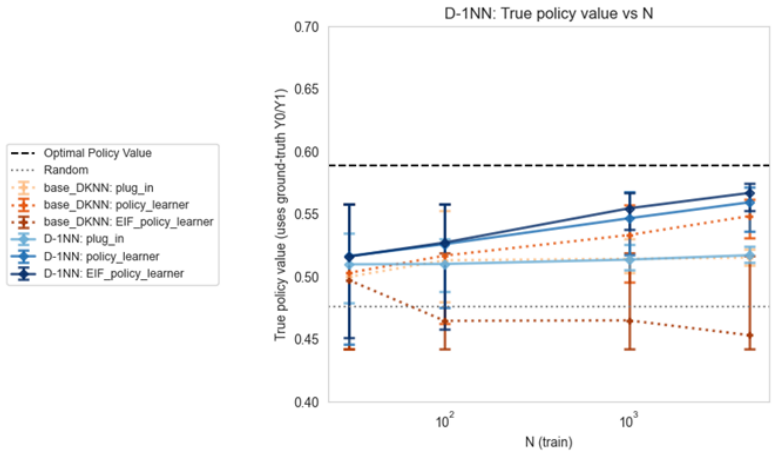


(d) DKNN: -0.5 (e) DKNN: 0 (f) DKNN: 0.5

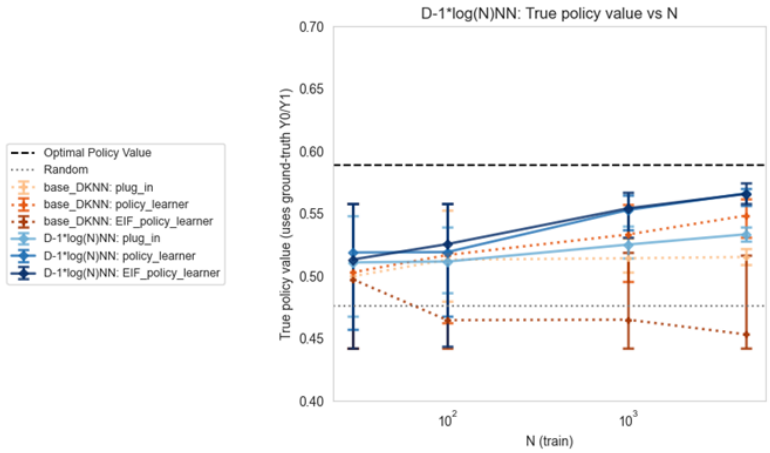


(g) DRF: -0.5 (h) DRF: 0 (i) DRF: 0.5

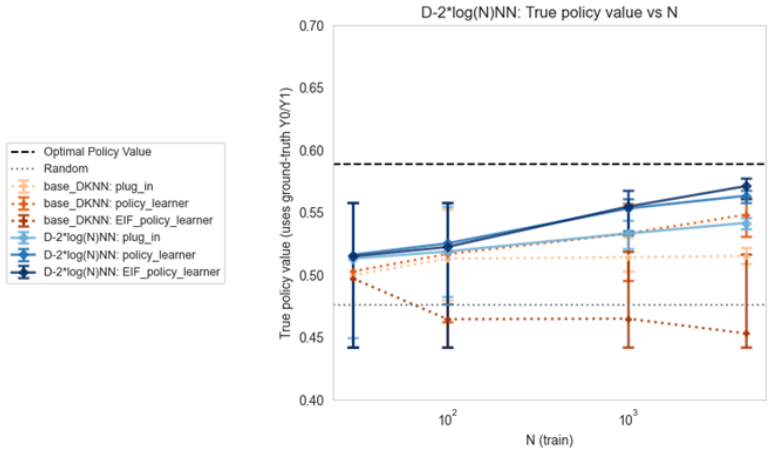
Figure 9: Policy Learning with correlation between potential outcomes in the semi-synthetic setting. (a–c) Linear Regression; (d–f) DKNN; (g–i) DRF. **Legends** as in Figure 1a.



(a) $k = 1$ neighbor

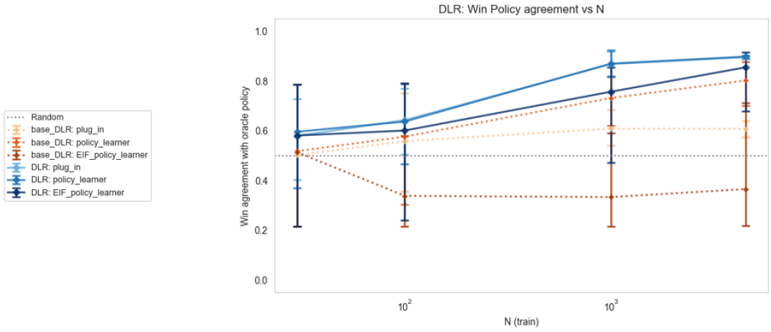


(b) $k = \log(n)$ neighbors

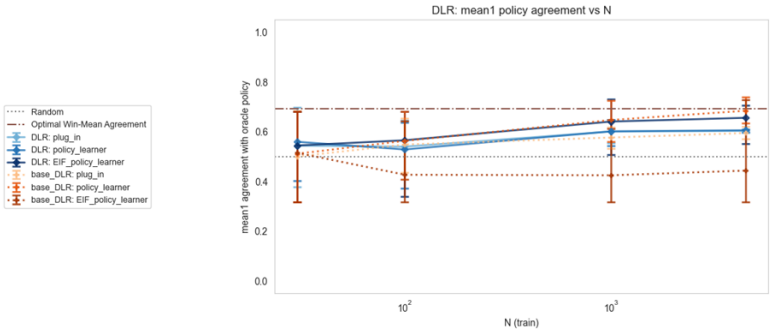


(c) $k = 2\log(n)$ neighbors

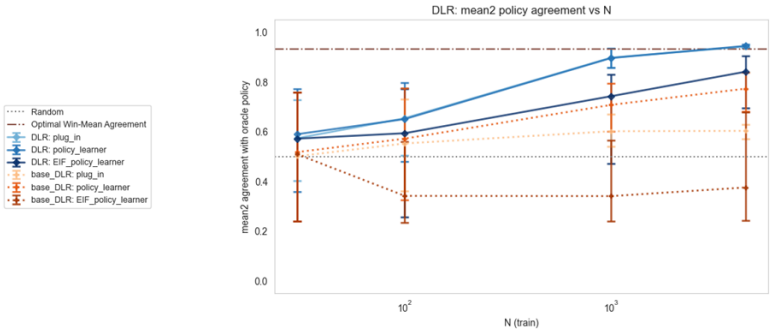
Figure 10: $k-$NN with $k = 1$, $k = \log(n)$ and $k = 2\log(n)$ on the semi-synthetic experiment without correlation



(a) Agreement with oracle policy optimizing by win



(b) Agreement with oracle policy optimizing by primary outcome



(c) Agreement with oracle policy optimizing by secondary outcome

Figure 11: Learned Policy's agreement with different oracle policies