CONSTRUCTING CONFIDENCE INTERVALS FOR AVERAGE TREATMENT EFFECTS FROM MULTIPLE OBSERVATIONAL DATASETS

Anonymous authors
Paper under double-blind review

ABSTRACT

Estimating confidence intervals (CIs) of the average treatment effects (ATE) from patient records is crucial to assess the effectiveness and safety of drugs. However, patient records typically come from different hospitals, thus raising the question of how multiple observational datasets can be effectively combined for this purpose. In our paper, we propose a new method that estimates the ATE from multiple observational datasets and provides valid CIs. Our method makes little assumptions about the observational datasets and is thus widely applicable in medical practice. The key idea of our method is that we leverage prediction-powered inferences and thereby essentially 'shrink' the CIs so that we offer more precise uncertainty quantification as compared to naïve approaches. We further prove the unbiasedness of our method and the validity of our CIs. We confirm our theoretical results through various numerical experiments. Finally, we provide an extension of our method for constructing CIs from combinations of experimental and observational datasets.

1 INTRODUCTION

Estimating the *average treatment effects (ATEs)* together with *confidence intervals (CIs)* is relevant in many fields such as medicine, where the ATE is used to assess the effectiveness and safety of drugs, including for drug approval (Glass et al., 2013; Feuerriegel et al., 2024). Nowadays, there is a growing interest in using observational datasets for this purpose, such as, for example, electronic health records (EHRs) and clinical registries (Johnson, 2016; Corrigan-Curay et al., 2018; Hong, 2021). Importantly, such observational datasets typically originate from different hospitals, different health providers, or even different countries (Colnet et al., 2023), thus raising the question of *how to construct CIs for ATE estimation from multiple observational datasets*.

Motivating example: During the COVID-19 pandemic, the effectiveness and safety of potential drugs and vaccines were often assessed from electronic health records that originated from different hospitals to rapidly generate new evidence with treatment guidelines (Tacconelli et al., 2022). For example, one study (Wong et al., 2024) estimated the effect of *nirmatrelvir/ritonavir* (also known under the commercial name "paxlovid") in patients with COVID-19 diagnosis on 28-day all-cause hospitalizations from data obtained through a retrospective, multi-center study. The study eventually reported not only the ATE but the corresponding CIs to allow for uncertainty quantification, which is standard in medicine (Kneib et al., 2023)

Existing works for estimating ATEs from *multiple* datasets can be loosely categorized based on **(a)** which datasets are used and **(b)** the underlying objective as follows (see Fig. 1): **(a)** The underlying patient data can come either from experimental datasets (i.e., randomized controlled trials; RCTs) and/or observational datasets (Feuerriegel et al., 2024). Both require tailored methods as the propensity score is known in RCTs but not in observational data and must thus be estimated (Rubin, 1974). We later focus on setting where the ATE is estimated from multiple observational datasets but we also provide an extension for combinations of RCT and observational datasets. **(b)** Much of the literature is focused on estimating ATEs from multiple datasets focuses on *point estimates* (Kallus et al., 2018; John et al., 2019; Yang & Ding, 2020; Guo et al., 2021; Hatt et al., 2022; Demirel et al., 2024), but ***not*** uncertainty quantification. Therefore, these methods do ***not*** provide valid CIs.

However, valid CIs are needed in medicine to ensure reliable decision-making, because of which the existing methods are *not* applicable for medical applications.

Our method: In this paper, we propose a novel method to *construct valid CIs for ATE estimation from multiple observational datasets*. Specifically, we consider a setting where we have one (potentially small) unbiased observational dataset $\mathcal{D}^1$ where we assume unconfoundedness (i.e., all confounders observed), and another large-scale observational dataset $\mathcal{D}^2$ where we additionally allow for unobserved confounders. Then, the key idea of our method is that we tailor prediction-powered inferences to our task so that we can essentially 'shrink' the CI and thus offer more precise uncertainty quantification as compared to naïve approach. We further present an extension of our method where we show to 'shrink' the CI in settings with a combination of RCT and observational data.

Why are naïve approaches precluded? One may think that one can simply concatenate both datasets to compute a pooled ATE, yet this is prohibited because the second dataset $\mathcal{D}^2$ may be confounded and, hence, the overall ATE estimate would be *biased*. A different, naïve approach is simply construct finite-sample CIs from the observational dataset $\mathcal{D}^1$ to obtain an unbiased ATE with valid CIs. Yet, the additional power of the second dataset $\mathcal{D}^2$ (e.g., the information about the treatment assignment and thus the propensity score) would be ignored, so that the CIs are *too conservative* (Aronow et al., 2021).

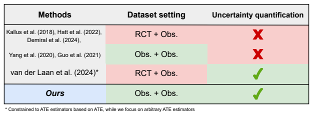
Intuition behind our method: The key, *non-trivial* challenge in our task is that even though the second dataset $\mathcal{D}^2$ is large and may help shrink the estimation variance, it may be confounded, which would lead to biases in the downstream estimation. Therefore, simply using the second dataset $\mathcal{D}^2$ directly for inference could lead to biased CATE estimates. To account for this, we derive a prediction-powered inference estimator where we decompose the variance of the population-level estimate of the CATE into two parts: one part comes from the estimation variance of the CATE on dataset $\mathcal{D}^2$, while the second part is due to the difference in estimators of ATEs across both datasets $\mathcal{D}^1$ and $\mathcal{D}^2$. The estimation variance of the first part can be significantly decreased with access to a large-scale dataset $\mathcal{D}^2$. Interestingly, the second part allows us to account for potential confounding bias in $\mathcal{D}^2$ and thus still derive valid CIs (→ Theorem 4.2)[1].

Our **main contributions** are three-fold:[2] (1) We propose a new method to construct CIs for the ATE from multiple observational datasets. We further extend our method to combinations of RCTs and observational datasets. (2) We prove that our method is a consistent estimator and gives valid CIs. (3) We perform experiments with medical data, demonstrating the effectiveness of our method.

2 RELATED WORK

In this section, we give an overview of three literature streams relevant to our work: (i) methods for constructing CIs for the ATE that solely rely on a *single* dataset; (ii) methods that estimate the ATE from *multiple* datasets; and (iii) prediction-powered inference.

Estimating CIs for the ATE: Several works focus on constructing confidence intervals (CIs) for the ATE in different settings (Bang & Robins, 2005; Laan & Rubin, 2006). One stream addresses asymptotically normal data, which typically results in $\sqrt{n}$-consistent, asymptotically unbiased, normally distributed estimators. For example, CIs are then constructed by adding and subtracting the product of the standard deviation and the conventional critical value of 1.96 from the standard normal distribution to yield 95% CIs (Hahn, 1998; Heckman et al., 1998; Winship & Morgan, 1999; Hirano et al., 2003; Chen et al., 2008). Another stream focuses on finite-sample settings, yet these works impose strong assumptions, such as that the data comes from an RCT (Aronow et al., 2021) or assume unconfoundedness with relaxed overlap assumptions (Armstrong & Kolesár, 2021). However, this stream focuses on ATE estimation from a single dataset, which is unlike our work.



Methods	Dataset setting	Uncertainty quantification
Kallus et al. (2018), Hatt et al. (2022), Demiral et al. (2024),	RCT + Obs.	✗
Yang et al. (2020), Guo et al. (2021)	Obs. + Obs.	✗
van der Laan et al. (2024)*	RCT + Obs.	✓
Ours	Obs. + Obs.	✓

* Constrained to ATE estimators based on ATE, while we focus on arbitrary ATE estimators

Figure 1: Key works aimed at ATE estimation from multiple datasets.

---

[1] We refer to Barnard (1949) and refer to a confidence interval a "valid" when the interval achieves its stated coverage probability. For example, a 95% confidence interval is valid if, under repeated sampling, it contains the true parameter value approximately 95% of the time. Validity ensures the interval accurately reflects the level of uncertainty about the estimate.

[2] Code is available via `https://anonymous.4open.science/r/causalppi-7BE5/`. Upon acceptance, we will move it to a public Repository.

ATE estimation from *multiple* datasets:[3] Existing methods can be grouped by **(a)** the underlying dataset setting and **(b)** the objective, that is, whether the focus is on point estimates vs. uncertainty quantification (see Fig. 1). We summarize both in the following. **(a)** Some methods focus on settings with RCT + observational data (e..g, Kallus et al., 2018; Chen et al., 2021; Hatt et al., 2022; Demirel et al., 2024).[4] Other methods focus on multiple observational data (e.g., Yang & Ding, 2020; Guo et al., 2021), which is our focus later. The former is typically easier because the propensity score is known, while, in the latter, the propensity score is not known but must be estimated to account for the covariate shift across treated and non-treated patients. **(b)** Most work focus on only point estimation (e..g, Kallus et al., 2018; Chen et al., 2021; Hatt et al., 2022; Yang & Ding, 2020; Guo et al., 2021; Demirel et al., 2024), but **not** uncertainty quantification. Hence, the output are **not** CIs, which is our objective.

Closest to our method is the work by van der Laan et al. (2024). Yet, there are crucial *differences*: (i) different ATE estimation process, (ii) different flexibility of leveraging $\mathcal{D}^2$. We discuss further differences in change to our new Appendix H.

Prediction-powered inference (PPI): Angelopoulos et al. (2023; 2024) proposed the PPI framework for performing valid statistical inference from a given dataset when the dataset is supplemented with predictions from a machine-learning model (a brief overview is in Sec. C). Several works have extended the original PPI framework (e.g., Zrnic & Candès, 2024; Fisch et al., 2024). Yet, PPI is *not* an 'off-the-shelf' framework; rather, the so-called rectified in PPI must be carefully derived for each statistical quantity of interest, which is *non-trivial*. So far, PPI was derived mostly for traditional statistical quantities (e.g., mean, median, quantile). For example, Demirel et al. (2024) derive PPI to generalize point estimates of causal effects from one population to a target population but *without* uncertainty quantification. However, to the best of our knowledge, there is **no** work that has tailored PPI to construct CIs in ATE estimation, which is our novelty.

Research gap: Existing methods for ATE estimation from multiple observational datasets focus on point estimates but *not* uncertainty quantification. To the best of our knowledge, we later derive the first method for constructing asymptotically valid CIs that focuses on this setting.

3 PROBLEM SETUP

We consider the standard setting for ATE estimation from observational data (e.g., Imbens, 2004; Rubin, 2006; Shalit et al., 2017; Hatt et al., 2022) but which we extend to multiple datasets.

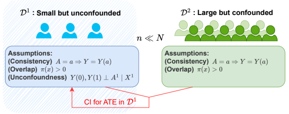

$\mathcal{D}^1$: Small but unconfounded
$\mathcal{D}^2$: Large but confounded
$n \ll N$
Assumptions:
(Consistency) $A = a \Rightarrow Y = Y(a)$
(Overlap) $\pi(x) > 0$
(Unconfoundness) $Y(0), Y(1) \perp A^1 \mid X^1$
Assumptions:
(Consistency) $A = a \Rightarrow Y = Y(a)$
(Overlap) $\pi(x) > 0$
CI for ATE in $\mathcal{D}^1$

Figure 2: **Setup** with two observational datasets and different assumptions on the underlying data-generating process.

Setting: We consider a setting with a small observational dataset $\mathcal{D}^1$ and a large-scale observational dataset $\mathcal{D}^2$ (see Fig. 2). We use $d \in D = \{1, 2\}$ to refer to the datasets. We write variables with superscript d to emphasize that variable $X^d \in \mathcal{D}^d$ for $\mathcal{D}^d \in \{\mathcal{D}^1, \mathcal{D}^2\}$. Without loss of generality, it is straightforward to extend our method to more than two datasets, simply by concatenating them into $\mathcal{D}^2$. Let n and N denote the size of the datasets, with $n \ll N$.

Both datasets have patient information about treatments, outcomes (e.g., tumor size, length of hospital stay, 30-day readmission risk), and covariates (e.g., the age or sex of a patient). Formally, both consist of assigned treatments $A_i^d \in \mathcal{A} = \{0, 1\}$, outcomes $Y_i^d \in \mathcal{Y} \subseteq \mathbb{R}$, and covariates $X_i^d \in \mathcal{X} \subseteq \mathbb{R}^q$ for dataset $d \in \{1, 2\}$ and with $i = 1, \ldots, n$ (for $d = 1$) and $i = 1, \ldots, N$ (for $d = 2$). Our setting is relevant for a variety of practical applications in medicine where electronic health records are collected from different environments, for example, from different hospitals or different countries.

---

[3] Several methods have also aimed at estimating heterogeneous treatment effects (HTEs) from multiple datasets (e.g., Johansson et al., 2018; Schweisthal et al., 2024). However, estimating HTEs is more challenging than estimating the ATE because of the variation across subpopulations and the larger risk of overlap violations. Importantly, using HTEs for computing ATEs is suboptimal, which is well-established in efficiency theory for ATE estimation (Kennedy, 2016) and would lead to so-called plug-in bias Curth & van der Schaar (2021). Hence, methods for HTE estimation are *orthogonal* to our work.

[4] There are specialized settings, yet which are different from ours. For example, some works estimate long-term outcomes by combining RCT and observational data (Athey et al., 2020; Ghassami et al., 2022; Imbens et al., 2024). Even others aim to increase the efficiency of trial analyses (Schuler et al., 2022; Liao et al., 2023).

We assume that data are sampled i.i.d from the same population $(X^d, A^d, Y^d) \sim \mathbb{P}$, meaning that patients come from the same population. We later also generalize our theory to settings with distribution shifts and finite populations in Appendix B.1 and Appendix B.2, respectively. Given that we focus on observational datasets, the treatment assignment rule may vary, and we thus define the dataset-specific propensity score via $\pi_d(x) = \mathbb{P}(A = 1 \mid X = x, D = d), d \in \{1, 2\}$. Formally, we assume that the propensity score may differ across the two datasets (i.e., $\pi_1 \neq \pi_2$). This is common in medical practice where different hospitals or countries have different treatment guidelines.

Target estimand: We adopt the potential outcomes framework (Neyman, 1923; Rubin, 1974) to formalize our causal inference task. Let $Y(a) \in \mathcal{Y}$ denote the potential outcome in the target population (i.e., where the small dataset is sampled from) for treatment intervention $A = a$. In this paper, we are interested in estimating the ATE given by $\tau = \mathbb{E}[Y(1) - Y(0)]$ in $\mathcal{D}^1$ and in constructing corresponding CIs for τ.

Assumptions: We make the following assumptions necessary for ATE identification and estimation. Of note, the following assumptions are standard in ATE estimation (Imbens, 2004; Rubin, 2006; Shalit et al., 2017). Here, we distinguish our assumptions for the small dataset $\mathcal{D}^1$ and the large $\mathcal{D}^2$.

Assumption 3.1. *For dataset $\mathcal{D}^1$, it holds: (i) (Consistency) $A = a \Rightarrow Y = Y(a)$; (ii) (Overlap) $0 < \pi(X) < 1, \forall X \in \mathcal{X}$; (iii) (Unconfoundedness) $Y(0), Y(1) \perp\!\!\!\perp A \mid X$.*

Assumption 3.2. *For dataset $\mathcal{D}^2$, it holds: (i) (Consistency) $A = a \Rightarrow Y = Y(a)$; (ii) (Overlap) $0 < \pi(X) < 1, \forall X \in \mathcal{X}$.*

The above assumptions are the *standard* assumptions for estimating ATEs from observational data and are widely used for any underlying estimation method (Imbens, 2004; Rubin, 2006; Shalit et al., 2017). Consistency holds as long as health information is accurately and systematically recorded. Overlap can be ensured through preprocessing (e.g., clipping). Unconfoundedness is plausible in digital health settings due to the growing availability of rich electronic health records.[5]

The above assumptions are consistent with the literature studying multiple observational datasets (Yang & Ding, 2020; Guo et al., 2021). Note that the assumptions for dataset $\mathcal{D}^2$ are weaker as compared to dataset $\mathcal{D}^1$. • For $\mathcal{D}^1$, we assume that there is no unobserved confounding but the propensity score is unknown. This is often the case in specialized medical facilities where patients receive close supervision and where thus all critical health measurements are reported, which is typically the case in cancer care (Castellanos et al., 2024) and in intensive care units (Johnson, 2016). Needless to say, our assumption is still considerably weaker than assuming an RCT because we allow that the treatment assignment mechanism varies greatly across subpopulations, is unknown, and must thus be estimated. Nevertheless, RCTs are a special case in which the RCT the propensity score π_1 is known. • For $\mathcal{D}^2$, we do *not* make the latter assumption but instead allow for unobserved confounding. This is often the case when data is recorded by general practitioners where the need for documentation is typically not as strictly enforced as in other medical facilities.

$\Rightarrow$ In sum, $\mathcal{D}^1$ would naturally lead to *unbiased* ATE estimation but suffers from a *large estimation variance* due to the small sample size. In contrast, $\mathcal{D}^2$ has a larger size and thus *more statistical power* but could lead to *biased* estimates *due to unobserved confounding*.

4 OUR METHOD FOR ATE ESTIMATION FROM MULTIPLE DATASETS

Overview. The general idea of our approach is shown in Fig. 3. Ⓐ **Measure of fit:** We first compute a measure of fit, m_θ, to estimate the ATE on the large, observational dataset $\mathcal{D}^2$. Here, we use a state-of-the-art method based on the DR-learner (Wager, 2024). We refer to the estimate as $\hat{\tau}_2$. Yet, $\hat{\tau}_2$ can be biased due to unobserved confounding because of which we later need to adjust for this via the rectifier. Ⓑ **Influence function estimation:** We compute the non-centered influence function score $\tilde{Y}_{\hat{\eta}}(x)$ for the observational dataset $\mathcal{D}^1$. This is later used in the rectifier. Ⓒ **Rectifier:** We compute the rectifier Δ_τ, which we use to adjust for the bias between datasets $\mathcal{D}^1$ and $\mathcal{D}^2$. This allows us transform the biased estimates $\hat{\tau}_2$ into unbiased estimates of the ATE in population. Ⓓ **Constructing CIs:** Eventually, this yields our CIs, $\mathcal{C}_\alpha^{\mathrm{PP}}$ for significance level α.

---

[5] Furthermore, advances in sensitivity analysis (Frauen et al., 2023; Oprescu et al., 2023) and partial identification (Duarte et al., 2023) offer complementary pathways to relax this assumption. In that sense, all existing works (see Fig. 1) for causal inference from multiple datasets make this assumption. In that sense, our work makes *weaker* assumptions that are more realistic as we allow for unobserved confounders in $\mathcal{D}^2$.

Why is the above task challenging? *First,* there is an information gap between different datasets. This means that different datasets come from different distributions, and we do not have any prior knowledge of the relationship between datasets. In particular, there can be a distribution shift due to various reasons, such as different populations $\mathcal{X}$ between both datasets, unobserved effect modifiers (i.e., variables that change the treatment effect even if they are no confounders), and different treatment assignment mechanisms because of which the propensity scores may be different across both datasets. We later account for this by proposing a rectifier that accounts for such distribution shifts through an AIPW-based estimation. Further, the propensity scores must be estimated, which introduces another source of uncertainty. *Second,* due to the fundamental problem of causal inference (Rubin, 1974), the ATEs are not directly observed but must be estimated, while considering the aforementioned distribution shift. Further, such estimates must be asymptotically valid, so that we can later derive CIs that are also asymptotically valid (→ see our Theorem 4.2).

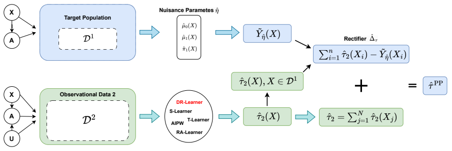

X
A
Target Population
$\mathcal{D}^1$
Nuisance Parametes $\hat{\eta}$
$\hat{\mu}_0(X)$
$\hat{\mu}_1(X)$
$\hat{\pi}_1(X)$
$\tilde{Y}_{\hat{\eta}}(X)$
Rectifier $\hat{\Delta}_\tau$
$\sum_{i=1}^n \hat{\tau}_2(X_i) - \tilde{Y}_{\hat{\eta}}(X_i)$
$\hat{\tau}_2(X), X \in \mathcal{D}^1$
+
=
$\hat{\tau}^{\mathrm{PP}}$
X
A
U
Observational Data 2
$\mathcal{D}^2$
DR-Learner
S-Learner
AIPW
T-Learner
RA-Learner
$\hat{\tau}_2(X)$
$\hat{\tau}_2 = \sum_{j=1}^N \hat{\tau}_2(X_j)$

Figure 3: **Overview of our method.** To construct CIs for the ATE, we leverage prediction-powered inferences: we decompose our task into computing a measure of fit (i.e., estimating the ATE on the large dataset $\mathcal{D}^2$ via the DR-learner, given by $\hat{\tau}_2$) and a rectifier $\hat{\Delta}_\tau$ (i.e., that measures the differences in ATE estimates across both datasets $\mathcal{D}^1$ and $\mathcal{D}^2$). However, finding a rectifier for our task is non-trivial and requires a careful derivation in order to ensure asymptotically valid CIs (→ our Theorem 4.2).

Below, we describe the steps Ⓐ–Ⓒ in detail. Pseuocode is in Algorithm 1.

Step Ⓐ: Measure of fit. The first step is to estimate the conditional average treatment effect (CATE) $\tau_2(x) = \mathbb{E}[Y(1) - Y(0) \mid X = x]$ on the large-scale dataset $\mathcal{D}^2$. Let $\hat{\tau}_2$ denote an arbitrary estimator (which may be biased due to, e.g., unobserved confounding). For example, we could choose the DR-learner due to its fast convergence rate and several favorable theoretical properties (Kennedy, 2023). Needless to say, one method is also applicable to other estimators without making any assumptions. Formally, we train $\hat{\tau}_2$ on $\mathcal{D}^2$. We then yield the measure of fit $\hat{\tau}_2 = \frac{1}{N}\sum_{j=1}^N \hat{\tau}_2(x_j)$.

Step Ⓑ: Influence function estimation. For our proposed rectifier, we later need the non-centered influence function (IF) score. Ideally, one would directly compute the difference in ATEs across both datasets for the rectifier, but this is impossible since the ATEs are not directly observed but rather need to be estimated. The estimation needs to be both valid and unbiased to later yield valid CIs. We observe that the average of the non-centered IF score of the AIPW estimator is an unbiased estimation of ATE and is asymptotically normally distributed. This is beneficial for two reasons: (i) we get an unbiased estimate, which allows us to later obtain an unbiased ATE in population, and (ii) the estimate is asymptotically normal so that we later can derive valid CIs.

Formally, the non-centered IF score for AIPW estimator (Robins & Rotnitzky, 1995) is given by

$$\tilde{Y}_{\hat{\eta}}(x_i) = \left(\frac{A}{\hat{\pi}(x_i)} - \frac{1-A}{1-\hat{\pi}(x_i)}\right) Y_i - \frac{A - \hat{\pi}(x_i)}{\hat{\pi}(x_i)\,(1-\hat{\pi}(x_i))}\left[(1-\hat{\pi}(x_i))\,\hat{\mu}_1(x_i) + \hat{\pi}(x_i)\hat{\mu}_0(x_i)\right], \tag{1}$$

where the nuisance parameters $\hat{\eta}(x) = (\hat{\mu}_0(x), \hat{\mu}_1(x), \hat{\pi}(x))$ are plug-in estimators from $\mathcal{D}^1$. Then, the AIPW estimator is $\hat{\tau}^{\mathrm{AIPW}} = \frac{1}{n}\sum_{i=1}^n Y_{\hat{\eta}}(x_i)$. We leverage the following result from the causal inference literature (Chernozhukov et al., 2018; Wager, 2024).

Remark 4.1 (follows from Wager (2024))**.** *Suppose we have consistent estimated nuisance functions* $\hat{\eta}(x) = (\hat{\mu}_0(x), \hat{\mu}_1(x), \hat{\pi}(x))$ *trained using sample splitting in splitted datasets with converge rate* $\mathcal{O}(n^{-\alpha_\mu})$ *and* $\mathcal{O}(n^{-\alpha_\pi})$*, i.e.,* $n^{-\alpha_\mu}(\hat{\mu}_w(X_i) - \mu_w(X_i)) \xrightarrow{p} 0$, $w = 0, 1$ *and* $n^{-\alpha_\pi}(1/\hat{\pi}(X_i) - 1/\pi(X_i)) \xrightarrow{p} 0$, *we the have that* $\hat{\tau}^{\mathrm{AIPW}}$ *is asymptotically normally distributed with* $\sqrt{n}\left(\hat{\tau}^{\mathrm{AIPW}} - \tau\right) \xrightarrow{d} \mathcal{N}\left(0, V^{\mathrm{AIPW}}\right)$, *where* $V^{\mathrm{AIPW}} = \mathrm{Var}\left[\mu_1(x) - \mu_0(x)\right] + \mathbb{E}\left[\left(A\frac{Y-\mu_1(x)}{\pi(x)}\right)^2\right] + \mathbb{E}\left[\left((1-A)\frac{Y-\mu_0(x)}{1-\pi(x)}\right)^2\right]$[6].

Hence, the above lemma allows us to estimate the ATE for $\mathcal{D}^1$ and construct the corresponding CI with non-centered influence function scores, which we then use in the rectifier to assess the bias between both datasets $\mathcal{D}^1$ and $\mathcal{D}^2$.

---

[6] The strong double robustness exists here. If we use the estimated nuisance functions that are both consistent and the RMSE of $\hat{\mu}(w)(x)$ and $\hat{e}(x)$ decays fast enough, then the AIPW estimation is asymptotically normal to the oracle ATE.

Step Ⓒ: Rectifier. We now introduce our proposed rectifier to quantify the difference in ATE across both datasets. Formally, we define the rectifier Δ_τ as the difference of $\hat{\tau}_{\text{AIPW}}$ and $\hat{\tau}_2$ on $\mathcal{D}^1$. For individual observations i, we write $\hat{\Delta}_i = \tilde{Y}_{\hat{\eta}}(x_i) - \hat{\tau}_2(x_i)$. Note that our rectifier is carefully tailored to our task, and is non-trivial because, due to the fundamental problem of causal inference Rubin (1974), the ATEs are never observed but we need to leverage the influence functions score in order to be able to compute a valid and unbiased estimate. Formally, we have

$$\begin{aligned}\hat{\Delta}_\tau = &\frac{1}{n}\sum_{i=1}^{n}\left[\tilde{Y}_{\hat{\eta}}(x_i) - \hat{\tau}_2(x_i)\right] = \frac{1}{n}\sum_{i=1}^{n}\Bigg[\left(\frac{A_i}{\hat{\pi}(x_i)} - \frac{1-A_i}{1-\hat{\pi}(x_i)}\right)Y_i \\ &- \frac{A_i - \hat{\pi}(x_i)}{\hat{\pi}(x_i)\left(1-\hat{\pi}(x_i)\right)}\left[\left(1-\hat{\pi}(x_i)\right)\hat{\mu}_1(x_i) + \hat{\pi}(x_i)\hat{\mu}_0(x_i)\right] - \hat{\tau}_2(x_i)\Bigg].\end{aligned} \tag{2}$$

Then, the prediction-powered estimation of ATE on $\mathcal{D}^1$ is computed via

$$\hat{\tau}^{\text{PP}} = \hat{\Delta}_\tau + \hat{\tau}_2 = \frac{1}{n}\sum_{i=1}^{n}\hat{\Delta}_i + \frac{1}{N}\sum_{j=1}^{N}\hat{\tau}_2(x_j) = \frac{1}{n}\sum_{i=1}^{n}\left[\tilde{Y}_{\hat{\eta}}(x_i) - \hat{\tau}_2(x_i)\right] + \frac{1}{N}\sum_{j=1}^{N}\hat{\tau}_2(x_j) \tag{3}$$

$$\begin{aligned}= &\frac{1}{N}\sum_{j=1}^{N}\hat{\tau}_2(x_j) + \frac{1}{n}\sum_{i=1}^{n}\Bigg[\left(\frac{A_i}{\hat{\pi}(x_i)} - \frac{1-A_i}{1-\hat{\pi}(x_i)}\right)Y_i \\ &- \frac{A_i - \hat{\pi}(x_i)}{\hat{\pi}(x_i)\left(1-\hat{\pi}(x_i)\right)}\left[\left(1-\hat{\pi}(x_i)\right)\hat{\mu}_1(x_i) + \hat{\pi}(x_i)\hat{\mu}_0(x_i)\right] - \hat{\tau}_2(x_i)\Bigg].\end{aligned} \tag{4}$$

Step Ⓓ: Constructing CIs. We now use the above PPI-based ATE estimate to construct our CIs. Let $\hat{\sigma}^2_{\tau_2}$ denote the empirical variance of $\hat{\tau}_2(x)$, and let $\hat{\sigma}^2_\Delta$ denote the empirical variance of $\hat{\Delta}_\tau$. Then, for significance level $\alpha \in (0,1)$, our prediction-powered confidence interval is

$$\mathcal{C}^{\text{PP}}_\alpha = \left(\hat{\tau}^{\text{PP}} \pm z_{1-\frac{\alpha}{2}}\sqrt{\frac{\hat{\sigma}^2_\Delta}{n} + \frac{\hat{\sigma}^2_{\tau_2}}{N}}\right), \tag{5}$$

where $\hat{\sigma}^2_\Delta = \frac{1}{n}\sum_{i=1}^{n}\left(\tilde{Y}_{\hat{\eta}}(x_i) - \hat{\tau}_2(x_i) - \hat{\Delta}_\tau\right)^2$, and $\hat{\sigma}^2_{\tau_2} = \frac{1}{N}\sum_{j=1}^{N}\left(\hat{\tau}_2(x_j) - \hat{\tau}_2\right)^2$. We show theoretically that $\mathcal{C}^{\text{PP}}_\alpha$ is asymptotically valid in Theorem 4.2.

Equation 5 has several implications for how our method 'shrinks' CIs. (i) The width of the CIs depends on the size of the different datasets (which we later evaluate empirically as part of our sensitivity analyses). Hence, the width shrinks with a larger dataset $\mathcal{D}^1$ and/or a larger dataset $\mathcal{D}^2$. (ii) The with of the CIs depends on the estimation variance $\hat{\sigma}^2_{\tau_2}$ from the dataset $\mathcal{D}^2$. This is desired because our method is particularly designed for using large-scale but confounded datasets $\mathcal{D}^2$, so that this term should shrink the CIs. (iii) The CIs further depend on the estimation variance of the rectifier $\hat{\sigma}^2_\Delta$. This term becomes smaller, the less confounding the observational dataset $\mathcal{D}^2$ has.

Theorem 4.2 (Validity of our prediction-powered CIs)**.** *Let $\mathcal{D}^1$ and $\mathcal{D}^2$ are sampled i.i.d. under the assumptions from above. Further suppose that we have consistent estimated nuisance functions $\hat{\eta}(X) = (\hat{\mu}_0(X), \hat{\mu}_1(X), \hat{\pi}(X))$ with converge rates $\mathcal{O}(n^{-\alpha_\mu})$, $\mathcal{O}(n^{-\alpha_\pi})$ and $\alpha_\mu + \alpha_\pi \geq 1/2$ and $\lim_{n,N\to\infty}\frac{n}{N} = p$ for some $p \in [0,1]$. Fix $\alpha \in (0,1)$, and let $\mathcal{C}^{\text{PP}}_\alpha = \left(\hat{\tau}^{\text{PP}} \pm z_{1-\alpha/2}\sqrt{\frac{\hat{\sigma}^2_\Delta}{n} + \frac{\hat{\sigma}^2_{\tau_2}}{N}}\right)$. Then, it holds that $\limsup_{n,N\to\infty} P(\tau \in \mathcal{C}^{\text{PP}}_\alpha) \geq 1-\alpha$.*

Proof. See Appendix A.2 where we leverage Lemma 4.1. □

Specifically, given that τ_2 is derived from a sum of independent random variables, the CLT ensures its asymptotic normality under standard regularity conditions, i.e., $\sqrt{n}(\hat{\tau}_2 - \mathbb{E}[\tau_2]) \xrightarrow{p} \mathcal{N}(0, \sigma^2_{\tau_2})$, as $n \to \infty$. Then our proposed estimator is a biased estimation of the oracle ATE, i.e., $\mathbb{E}[\hat{\tau}^{\text{PP}}] = \tau$ and the key focus of our analysis lies in the asymptotic normality of $\hat{\tau}_1$ in the observational unconfounded dataset.

Algorithm 1 Prediction-powered CIs for ATE estimation from multiple observational datasets

```
Input: small dataset $\mathcal{D}^1 = \left(x^1, A^1, Y^1\right)$, large dataset $\mathcal{D}^2 = \left(x^2, A^2, Y^2\right)$, significance level $\alpha \in (0,1)$
$\hat{\tau}_2(x) \leftarrow$ estimate CATE estimator from $\mathcal{D}^2$ and $\tilde{Y}_{\hat{\eta}}(x) \leftarrow$ estimate non-centered influential function score from $\mathcal{D}^1$
$\hat{\Delta}_i \leftarrow \tilde{Y}_{\hat{\eta}}(x_i) - \hat{\tau}_2(x_i)$
$\hat{\tau}_2 \leftarrow \frac{1}{N}\sum_{i=1}^{N} \hat{\tau}_2(x_i)$, and $\hat{\Delta}_\tau \leftarrow \frac{1}{n}\sum_{i=1}^{n} \hat{\Delta}_i$
$\hat{\tau}^{\mathrm{PP}} \leftarrow \hat{\tau}_2 - \hat{\Delta}_\tau$    ▷ prediction-powered estimator
$\hat{\sigma}^2_{\tau_2} \leftarrow \frac{1}{N}\sum_{i=1}^{N} (\hat{\tau}_2(x_i) - \hat{\tau}_2)^2$    ▷ empirical variance of CATE estimation in $\mathcal{D}^2$
$\hat{\sigma}^2_{\Delta} \leftarrow \frac{1}{n}\sum_{i=1}^{n} \left(\hat{\Delta}_i - \hat{\Delta}_\tau\right)^2$    ▷ empirical variance of rectifier in $\mathcal{D}^1$
$w_\alpha \leftarrow z_{1-\frac{\alpha}{2}} \sqrt{\frac{\hat{\sigma}^2_{\Delta}}{n} + \frac{\hat{\sigma}^2_{\tau_2}}{N}}$    ▷ normal approximation
Output: prediction-powered confidence interval $\mathcal{C}^{\mathrm{PP}}_{\alpha} = \left(\hat{\tau}^{\mathrm{PP}} \pm w_\alpha\right)$
```

Why is our method better than using the unconfounded dataset only? As shown in Equation 5, the width of our proposed CIs are mainly determined by the variance term, $\sqrt{\frac{\hat{\sigma}^2_{\Delta}}{n} + \frac{\hat{\sigma}^2_{\tau_2}}{N}}$. When the $\hat{\tau}_2$ are sufficiently accurate, the rectifier is almost equal to zero, i.e., $\hat{\Delta} \approx 0$. Then, the variance of the rectifier is significantly smaller than the variance of estimated non-centered IF scores, i.e., $\hat{\sigma}^2_{\Delta} \leq \hat{\sigma}^2_{\tau_2}$. Given the large size of $\mathcal{D}_2$, the variance of the estimated conditional treatment effect goes to zero, since the estimated variance should be divided by the sample size of $\mathcal{D}^2$, i.e., N. As a result, the variance (and thus the CI width) is smaller when using our method than when using only the unconfounded dataset, which means that our CIs are more narrow than the naïve baseline.

The above theorem is crucial because it ensures that our PPI-based CIs are asymptotically valid. Further, note that the above theorem is our contribution: it does *not* directly follow from the PPI-based framework. Rather, we must carefully leverage theoretical guarantees for the estimand of interest and the chosen rectifier, which is one of our contributions.

5 EXTENSION OF OUR METHOD FOR RCT + OBSERVATIONAL DATASETS

We now extend our PPI-based method to combinations of RCT+observational data. Using an RCT dataset is a special case of $\mathcal{D}^1$. As a result, the propensity score is known, which simplifies the underlying task. Yet, the information gap between the datasets remains in that both come from different distributions (e.g., different populations X, different effect modifiers, etc.).

A straightforward way would be to apply our AIPTW-based method directly with the known propensity. However, this may have disadvantages as it still requires the estimation of nuisance functions (response functions). Below, we describe the alternative method based on the inverse-propensity weighting (IPW) estimator, which necessary changes for the steps Ⓐ–Ⓒ. Note that we do longer need the influence function estimation because the propensity score is known so that we can directly estimate τ_1 via the IPW estimator.

Step Ⓐ: Measure of fit. This steps compute the CATEs analogous to the above by training $\hat{\tau}_2$ on $\mathcal{D}^2$. We thus yield the measure of fit $\hat{\tau}_2 = \frac{1}{N}\sum_{j=1}^{N} \hat{\tau}_2(x_j)$.

Step Ⓑ: IPW estimator. Given the RCT dataset ($\mathcal{D}^1$) and known propensity score, we can compute the inverse propensity weighted estimation of ATE via

$$\hat{\tau}_1 = \frac{1}{n}\sum_{i=1}^{n} \tilde{Y}_\pi(x_i) = \frac{1}{n}\sum_{i=1}^{n} \left(\frac{A_i Y_i}{\pi(x_i)} - \frac{(1-A_i)Y_i}{\pi(x_i)} \right), \tag{6}$$

which we later use in the rectifier (instead of the influence function score as in our method for multiple observational datasets).

Remark 5.1. [7] *Let $\hat{\tau}_1$ denote the IPW estimator for the ATE of the RCT dataset ($\mathcal{D}^1$), $\hat{\tau}_1$ is asymptotically normally distributed, i.e., $\sqrt{n}\,(\hat{\tau}_1 - \tau) \xrightarrow{d} \mathcal{N}\left(0, \hat{\sigma}_1^2\right)$, where $\hat{\tau}_1 = \frac{1}{n}\sum_{i=1}^{n} \tilde{Y}_\pi(x_i)$, $\hat{\sigma}_1^2 = \frac{1}{n}\sum_{i=1}^{n} \left(\tilde{Y}_\pi(x_i) - \hat{\tau}_1\right)^2$.*

The above Lemma 5.1 ensures that the estimate is asymptotically normal, which allows us to later obtain valid CIs.

Step Ⓒ: Rectifier. We now introduce our rectifier Δ_τ, which denotes the bias of two CATE estimators on $\mathcal{D}^1$. For this, let $\hat{\tau}_1(x)$ and $\hat{\tau}_2(x)$ be separately trained CATE estimators on $\mathcal{D}^1$ and $\mathcal{D}^2$, respectively. Our rectifiers then given by $\Delta_\tau = \mathbb{E}\left[\tau_1(x) - \tau_2(x)\right]$.

---

[7] The proof is standard and follows from, e.g., Wager (2024). We nevertheless provide a step-by-step variants in Appendix A for interested readers.

Then, the prediction-powered estimate of the ATE on $\mathcal{D}^1$ is computed via

$$\hat{\tau}^{\mathrm{PP}} = \frac{1}{N}\sum_{j=1}^{N}\hat{\tau}_2(x_j) + \frac{1}{n}\sum_{i=1}^{n}\hat{\Delta}_i = \frac{1}{N}\sum_{i=1}^{N}\hat{\tau}_2(x_i) + \frac{1}{n}\sum_{j=1}^{n}\tilde{Y}_\pi(x_i) - \hat{\tau}_2(x_j). \tag{7}$$

Step Ⓓ: **Constructing CIs.** We now use the above PPI-based ATE estimate to construct our CIs. Let $\hat{\sigma}^2_{\tau_2}$ denote the empirical variance of $\hat{\tau}_2(\mathbf{x})$, and let $\hat{\sigma}^2_{\Delta}$ denotes empirical variance of rectifier. Then, for significance level $\alpha \in (0,1)$, the our prediction-powered CI is given by

$$\mathcal{C}^{\mathrm{PP}}_{\alpha} = \left(\hat{\tau}^{\mathrm{PP}} \pm z_{1-\frac{\alpha}{2}}\sqrt{\frac{\hat{\sigma}^2_{\Delta}}{n} + \frac{\hat{\sigma}^2_{\tau_2}}{N}}\right), \tag{8}$$

where $\hat{\Delta}_\tau = \frac{1}{n}\sum_{i=1}^{n}\hat{\Delta}_i$, $\hat{\sigma}^2_{\Delta} = \frac{1}{n}\sum_{i=1}^{n}\left(\hat{\Delta}_i - \hat{\Delta}_\tau\right)$, and $\hat{\tau}_2 = \frac{1}{N}\sum_{j=1}^{N}\hat{\tau}_2(x_j)$, $\hat{\sigma}^2_{\tau_2} = \frac{1}{N}\sum_{j=1}^{N}\left(\hat{\tau}_2(x_j) - \hat{\tau}_2\right)$. The following theorem establishes that our above prediction-powered CI is valid.

Theorem 5.2 (Validity of our prediction-powered CIs)**.** *Let* $\mathcal{D}^1$ *and* $\mathcal{D}^2$ *are sampled i.i.d. under the assumptions from above, and* $\lim_{n,N\to\infty}\frac{n}{N} = p$ *for some* $p \in [0,1]$. *Then, the prediction-powered confidence interval has valid coverage:* $\liminf_{n,N\to\infty} P\left(\tau \in \mathcal{C}^{\mathrm{PP}}_{\alpha}\right) \geq 1-\alpha$.

Proof. See Appendix A.2 where we leverage Lemma 5.1. ☐

6 EXPERIMENTS

We now evaluate the effectiveness of our proposed method by examining the faithfulness and width of the constructed CIs. To this end, we follow prior research and perform experiments with both synthetic and real-world medical data (e.g., Schröder et al., 2024; Schweisthal et al., 2024). Synthetic data has the advantage that we have access to the ground-truth CATEs and thereby can make comparisons against oracle estimates. Further, real-world medical data allows us to demonstrate both the applicability and relevance of our method in practice.

6.1 SYNTHETIC DATA

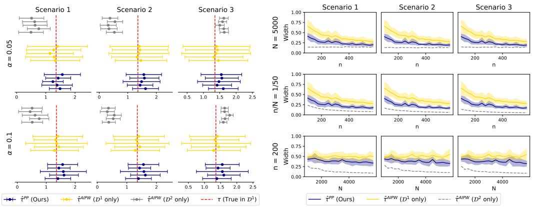


Figure 4: **Performance for synthetic dataset.** *Left:* We show the estimated CIs for five seeds. The red line is the oracle ATE. Ideally, the CIs should be narrow but still overlap with the oracle ATE. *Right:* Shows in the width of the CIs averaged over five different seeds ($\alpha = 0.05$). Here, we vary the size of the different datasets given by n ($\mathcal{D}^1$ only) and N ($\mathcal{D}^2$). Note that $\hat{\tau}^{\mathrm{AIPW}}$ ($\mathcal{D}^2$ only) is shown in intentionally shown in gray: it is *not* faithful as seen in the left plot and therefore *not* a valid baseline. ⇒ Our method yields faithful CIs, and it performs better in shrinking the width of CIs as desired.

Data: Inspired by Demirel et al. (2024), we simulate samples from a data-generating process with a confounder $x \in [-1,1]$ and a unobserved confounder $u \in U \subseteq [-1,1]$, a binary treatment $A \in \{0,1\}$, and a real-valued outcome $Y \in \mathbb{R}$. We generate the potential outcomes Y^d, $d \in \{1,2\}$ conditioned on x and u by sampling from a Gaussian process $\mathcal{GP} : [-1,1]^2 \to \mathbb{R}$ with mean function $m(x,u) = 0$ and kernel function $k\left((x,u),(x',u')\right)$. We choose a composite kernel by adding a squared-exponential (SE) kernel to model the *local* variation and a linear kernel to model the trends in outcome. As a result, we have $k\left((x,u),(x',u')\right) = \alpha_x x x' + \alpha_u u u' + \exp\left[-\frac{(x-x')}{2l_x^2} - \frac{(u-u')}{2l_u^2}\right]$, with configuration parameters $\theta = \{\alpha_x, \alpha_u, l_x, l_u\} \in \mathbb{R}^4_+$. We can simulate different confounding strengths by varying the value of θ.

We generate the covariates of observational datasets $\mathcal{D}^1$ and $\mathcal{D}^2$ by sampling $x_i, u_i \sim \mathrm{uniform}[-1,1]$ independently. For each patient, treatments assignments are sampled via $A_i \sim \mathrm{Bernoulli}(P(A=1 \mid x_i, u_i))$, where the probability of treatment assignment is generated similarly to Equation 6.1 via a logit function

$L_\pi(x, u)$ sampled from $\mathcal{GP}_{\theta_\pi}(x, u)$, where $\theta_\pi = \{\alpha_x^\pi, \alpha_u^\pi, l_x^\pi, l_u^\pi\}$. A larger value of α_u and a smaller value of l_u implies stronger confounding. The observed outcomes are computed via $Y = (1 - A) \cdot \mathcal{GP}_{\theta_0}(x, u) + A \cdot \mathcal{GP}_{\theta_1}(x, u)$.

We generate $n = 200$ ($\mathcal{D}^1$) and $N = 5000$ ($\mathcal{D}^2$) samples. For $\mathcal{D}^1$, we set $\alpha_u = 0$ and $l_u = 10^6$ to prevent confounding. For $\mathcal{D}^2$, we use different values of θ to generate different in confounding scenarios. We consider the following scenarios: • **Scenario 1:** little confounding (with 5.89). • **Scenario 2:** medium confounding (with 6.12). • **Scenario 3:** heavy confounding (with 7.69). Further details about the setting of θ are given in Appendix F.1. Altogether, we generate over 60 different datasets under varying configurations for evaluations below.

Baselines: We compare our PPI-based method $\hat{\tau}^{\text{PP}}$ for constructing CIs against the following baseline: **(1)** we estimate the ATE via the AIPW estimator $\hat{\tau}^{\text{AIPW}}$ only on the small dataset, named $\hat{\tau}^{\text{AIPW}}$ ($\mathcal{D}^1$ only) which is the **naïve baseline**; **(2)** we estimate the ATE via the AIPW estimator on the large, confounded dataset, named $\hat{\tau}^{\text{AIPW}}$ ($\mathcal{D}^2$ only); and **(3)** we report the true value for τ in $\mathcal{D}^1$.

Main results: Fig. 4 (left). We observe the following: **(1)** The CIs from our method overlap with the oracle ATE (in red), which shows that our method is faithful. **(2)** In contrast, the baseline $\hat{\tau}^{\text{AIPW}}$ ($\mathcal{D}^2$ only) rarely covers the oracle ATE and is thus *not* faithful. This can be expected since the dataset computes the CIs based on the confounded dataset and, hence, yields at biased estimates. The un-faithfulness becomes especially evident in Scenario 3 where data under large confounding is generated. **(3)** Our method generates CIs that are more narrow as compared to the naïve baseline $\hat{\tau}^{\text{AIPW}}$ ($\mathcal{D}^1$ only) and therefore clearly advantageous. For example, in the left plot, our CIs are, on average, smaller by 49.99% (Scenario 1), 55.37% (Scenario 2), and 55.35% (Scenario 3). ⇒ ***Takeaway:*** *Our PPI-based method yields faithful CIs but where the width of the CIs is clearly shorter. Hence, our method performs the best.*

Sensitivity to dataset size: Fig. 4 (right) compares the sensitivity to different dataset sizes. **(1)** Our method generates CIs that are again more narrow and therefore superior. We observe that our method outperforms the naïve baseline in all scenarios and across all dataset sizes. In other words, our method performs better in terms of widths of CIs than the naïve baseline. **(2)** The advantages of our method become pronounced for setting where $N \gg n$ as expected (see right plot, top row).

6.2 MEDICAL DATA

Dataset: We now provide a case study where demonstrate the applicability of our method to real-world medical datasets. We chose two common datasets: the **MIMIC-III** dataset (Johnson, 2016) and a Brazilian **COVID-19** dataset (Baqui et al., 2020). • **MIMIC-III** contains health records from patients admitted to intensive care units at large hospitals. We aim to estimate the average red blood cell count of all patients after being treated with mechanical ventilation. Our estimation is based on 8 confounders from medical practice (e.g., respiratory rate, hematocrit). • The **COVID-19** dataset contains health records of hospitalizations in Brazil across different regions and from patients with different socio-economic backgrounds. We are interested in predicting the effect of comorbidities on the mortality of COVID-19 patients. We created two different splits of the original dataset into $\mathcal{D}^1$ and $\mathcal{D}^2$: (i) once we split by regions of the hospitals in Brazil (i.e., North and Central-South) and (ii) once by ethnicity of participants (i.e., White and others). Further details are in Appendix F.2.

Table 1: **Results for different medical datasets.** We report the RMSE of the ATE estimator and the width of the CIs. The results for $\hat{\tau}^{\text{AIPW}}$ ($\mathcal{D}^2$ only) are shown in gray because the estimator is *not* faithful and therefore also *not* a viable baseline. Reported is the average performance over 5 random seeds.

Dataset	MIMIC-III		COVID-19 (by region)		COVID-19 (by ethnicity)	
	RMSE	Width	RMSE	Width	RMSE	Width
$\hat{\tau}^{\text{AIPW}}$ ($\mathcal{D}^1$ only)	**0.057**	0.077	7.591	8.479	39.970	0.081
$\hat{\tau}^{\text{AIPW}}$ ($\mathcal{D}^2$ only)	0.058	0.003	17.125	0.311	39.999	0.004
$\hat{\tau}^{\text{PP}}$ **(Ours)**	**0.057**	**0.023**	**7.131**	**2.341**	**39.968**	**0.026**

Smaller is better. Best value in bold.

Results: The results are in Table 1. We again compare the CIs of our estimator against the baselines from above. We further report the root means squared error for the factual outcomes. We find: **(1)** Our method achieves the smallest RMSE, which indicates that underlying patterns in the are well captured. **(2)** Our method obtains the smallest, yet valid CIs. Compared to $\hat{\tau}^{\text{AIPW}}$ in $\mathcal{D}^1$, our methods leads a ∼ 3.5x reduction in the width of CIs. **(3)** $\hat{\tau}^{\text{AIPW}}$ in $\mathcal{D}^2$ is known to be biased. This explains why the RMSE is sometimes considerably larger than the RMSE for the other methods, which again corroborates our findings that $\hat{\tau}^{\text{AIPW}}$ is not faithful. ⇒ ***Takeaway:*** *Our PPI-based method is effective for medical data.*

6.3 RESULTS FOR COMBINATIONS OF RCT+OBSERVATIONAL DATA

Data: We use the same data-generating process from above. However, we now mimic an RCT for $\mathcal{D}^1$ by setting the unobserved confounder U to zero and corresponding $\alpha = 0$ and $l_u = 10^6$. Details are in Appendix F.1.

Baselines: We now report our method based on IPW (instead of AIPW). We additionally implement the method by van der Laan et al. (2024), which allows to estimate the ATE from both datasets. We refer to this method by $\hat{\tau}^{\mathrm{ATMLE}}$. However, we note that the method is often unstable: their method involves a matrix inversion, yet where the matrix is often singular, so that no CIs can be computed (see Appendix H for a detailed explanation). This later explains the fairly noisy performance of the baseline. For a fair comparison, we simply set the output in these cases to $\mathcal{D}^1$.

Results: Fig. 5 (left) shows the results. We find: **(1)** The CIs from our method cover the oracle ATE (in red), which shows that our method is faithful for the RCT+observational dataset. **(2)** $\hat{\tau}^{\mathrm{ATMLE}}$ is faithful in the settings with little and medium confounding (Scenarios 2 and 3), but it fails in Scenario 3 where it is *not* faithful. **(3)** Our method generates CIs that are consistently more narrow compared to the baselines. ⇒ ***Takeaway:** Our method performs best.*

Sensitivity to dataset size: In Fig. 5 (right), we analyze the role of dataset sizes. The results confirm our findings from above: Compared to $\hat{\tau}^{\mathrm{ATMLE}}$, our method is much more stable. Further, our method generates CIs that are consistently more narrow and thus superior.

6.4 ADDITIONAL EXPERIMENTS

We provide further experiments to corroborate our above takeaways in (Appendix. G). • **Variations of our method:** (1) We performed experiments where we instantiated our method using **neural networks** as regression models for estimating nuisance parameters in AIPW to offer more flexibility in learning representation of the covariate space (see Appendix G.1). (2) We performed experiments with **XGBoost** to show the applicability of our method to underlying base models for estimating nuisance parameters in AIPW (see Appendix G.2). Here, our method based on neural networks performs best. • **Different settings:** (3) We varied the size of the covariate space to demonstrate the effectiveness of our method in settings with a **high-dimensional covariate space** (see Appendix G.3). (4) We varied the **covariate dependence** by increasing the collinearity in input space (see Appendix G.4). (5) We varied the **strength of confounding** in $\mathcal{D}^1$ (see Appendix G.5). We found that our method performs best across settings. • **Robustness/refutation checks:** (6) Oftentimes, estimates of treatment effects in RCT settings benefit when the propensity score is estimated (Su et al., 2023; Cai & van der Laan, 2019). Motivated by this, we applied our AIPW-based method to combinations of RCT and observational datasets (see Appendix G.6). Here, we find that we can improve the CI width further using our AIPW-based method. (7) We performed a refutation check in which we applied A-TMLE to combinations of two observational datasets (see Appendix G.7) but remind that this violates the assumptions of A-TMLE. As expected, A-TMLE underperforms, and our method remains clearly superior. (8) We expanded the sample size of $\mathcal{D}^1$ from 100 to 2500, to further assess the role of the size of $\mathcal{D}^1$ and the robustness of our method regarding the size of $\mathcal{D}^1$. We found that our method shows clear margin and results are as expected.

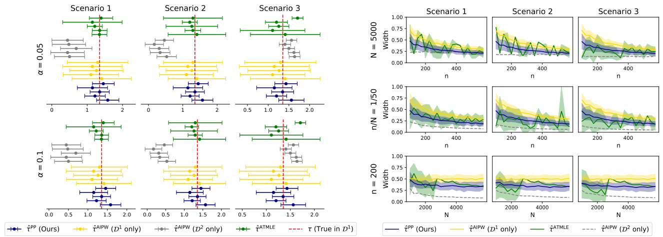


Figure 5: **Performance for synthetic dataset.** *Left:* We show the estimated CIs for five different seeds in RCT and observational datasets. *Right:* We show the width of the CIs averaged over five different seeds ($\alpha = 0.05$). ⇒ Our method is both stable and leads to CIs that are faithful and narrow, as desired.

7 DISCUSSION

Relevance: In this paper, we developed a new method for ATE estimation from multiple observational datasets. Our method is highly relevant to medical practice where it helps to assess the effectiveness and safety of drugs, and, to this end, we perform rigorous uncertainty quantification by deriving and reporting valid CIs. **Limitations & future work:** One improvement is extending our method to other estimands like the CATE or to causal survival analysis. Future research may combine our method with the pre-trained large language model (LLMs) or develop tailored neural network architectures on top of our method for text-based representations. However, as with any method in causal inference, the assumptions must be carefully assessed to ensure a safe and reliable use.

Notation

$\mathcal{S}$	labeled dataset
$\tilde{\mathcal{S}}$	unlabeled dataset
$\mathcal{D}^1$	Small dataset
$\mathcal{D}^2$	Large observational dataset
X, X_i	Confounders and confounders of i-th individual
A, A_i	Treatment and treatment of i-th individual
Y, Y_i	Real-valued outcome and outcome of i individual
$\pi_d(X)$	dataset-specific propensity score function
τ	Average treatment effect
$\hat{\tau}_1$	Estimated conditional average treatment effect
$\hat{\sigma}_1^2$	Empirical variance of conditional average treatment effect
$\hat{\tau}^{\mathrm{PP}}$	Prediction-powered ATE estimation
$\hat{\Delta}_\tau, \hat{\Delta}_i$	Mean of rectifier and rectifier of i-th inidividual
$\hat{\sigma}_\Delta^2$	Empirical variance of rectifier
$\hat{\tau}_2, \hat{\tau}_2(X_i)$	Mean of CATE in $\mathcal{D}^2$ and estimated CATE of i-th inidividual
$\hat{\sigma}_{\tau_2}^2$	Empirical variance of CATE in $\mathcal{D}^2$
$\hat{\eta}$	Nuisance parameters
$\hat{\mu}_0(X), \hat{\mu}_1(X)$	Regression function of average potential outcome
$\hat{\pi}(X)$	Estimated propensity score function
$\tilde{Y}_{\hat{\eta}}(X)$	Non-centered influence function score
α	Significance level
$z_{1-\alpha/2}$	The upper $1-\alpha/2$ quantile of the normal distribution
$\mathcal{C}_\alpha^{\mathrm{PP}}$	Prediction-powered $(1-\alpha)\%$ confidence interval
$\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$	Gaussian distribution with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$

REFERENCES

Anastasios N. Angelopoulos, Stephen Bates, Clara Fannjiang, Michael I. Jordan, and Tijana Zrnic. Prediction-powered inference. *Science*, 382(6671):669–674, 11 2023.

Anastasios N Angelopoulos, Stephen Bates, Adam Fisch, Lihua Lei, and Tal Schuster. Conformal Risk Control, 2024.

Timothy B Armstrong and Michal Kolesár. Finite-Sample Optimal Estimation and Inference on Average Treatment Effects Under Unconfoundedness. *Econometrica*, 2021.

P. M. Aronow, James M. Robins, Theo Saarinen, Fredrik Sävje, and Jasjeet Sekhon. Nonparametric identification is not enough, but randomized controlled trials are, 9 2021. arXiv:2108.11342.

Susan Athey, Raj Chetty, and Guido Imbens. Combining Experimental and Observational Data to Estimate Treatment Effects on Long Term Outcomes, 6 2020. arXiv:2006.09676.

Heejung Bang and James M. Robins. Doubly Robust Estimation in Missing Data and Causal Inference Models. *Biometrics*, 61(4):962–973, 2005. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1541-0420.2005.00377.x.

Pedro Baqui, Ioana Bica, Valerio Marra, Ari Ercole, and Mihaela van der Schaar. Ethnic and regional variations in hospital mortality from COVID-19 in Brazil: a cross-sectional observational study. *The Lancet Global Health*, 8(8):e1018–e1026, 8 2020.

G. A. Barnard. Statistical Inference, 1949.

Weixin Cai and Mark J. van der Laan. One-step targeted maximum likelihood for time-to-event outcomes, 2019.

Emily H. Castellanos, Brett K. Wittmershaus, and Sheenu Chandwani. Raising the Bar for Real-World Data in Oncology: Approaches to Quality Across Multiple Dimensions. *JCO Clinical Cancer Informatics*, 8:e2300046, 1 2024.

Shuxiao Chen, Bo Zhang, and Ting Ye. Minimax rates and adaptivity in combining experimental and observational data. *arXiv preprint arXiv:2109.10522*, 2021.

Xiaohong Chen, Han Hong, and Alessandro Tarozzi. Semiparametric efficiency in GMM models with auxiliary data. *The Annals of Statistics*, 36(2):808–843, 4 2008.

Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1):C1–C68, 2 2018.

Bénédicte Colnet, Imke Mayer, Guanhua Chen, Awa Dieng, Ruohong Li, Gaël Varoquaux, Jean-Philippe Vert, Julie Josse, and Shu Yang. Causal inference methods for combining randomized trials and observational studies: A review, 2023.

Jacqueline Corrigan-Curay, Leonard Sacks, and Janet Woodcock. Real-World Evidence and Real-World Data for Evaluating Drug Safety and Effectiveness. *JAMA*, 320(9):867–868, 9 2018.

Alicia Curth and Mihaela van der Schaar. Nonparametric Estimation of Heterogeneous Treatment Effects: From Theory to Learning Algorithms, 2 2021. arXiv:2101.10943.

Ilker Demirel, Ahmed Alaa, Anthony Philippakis, and David Sontag. Prediction-powered Generalization of Causal Inferences, 6 2024. arXiv:2406.02873.

Guilherme Duarte, Noam Finkelstein, Dean Knox, Jonathan Mummolo, and Ilya Shpitser. An automated approach to causal inference in discrete settings, 2023.

Stefan Feuerriegel, Dennis Frauen, Valentyn Melnychuk, Jonas Schweisthal, Konstantin Hess, Alicia Curth, Stefan Bauer, Niki Kilbertus, Isaac S. Kohane, and Mihaela van der Schaar. Causal machine learning for predicting treatment outcomes. *Nature Medicine*, 30(4):958–968, 4 2024.

Adam Fisch, Joshua Maynez, R. Alex Hofer, Bhuwan Dhingra, Amir Globerson, and William W. Cohen. Stratified Prediction-Powered Inference for Hybrid Language Model Evaluation, 6 2024. arXiv:2406.04291.

Dennis Frauen, Valentyn Melnychuk, and Stefan Feuerriegel. Sharp bounds for generalized causal sensitivity analysis. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 40556–40586. Curran Associates, Inc., 2023.

AmirEmad Ghassami, Alan Yang, David Richardson, Ilya Shpitser, and Eric Tchetgen Tchetgen. Combining Experimental and Observational Data for Identification and Estimation of Long-Term Causal Effects, 4 2022. arXiv:2201.10743.

Thomas A. Glass, Steven N. Goodman, Miguel A. Hernán, and Jonathan M. Samet. Causal inference in public health. *Annual Review of Public Health*, 34:61–75, 2013.

Wenshuo Guo, Serena Wang, Peng Ding, Yixin Wang, and Michael I Jordan. Multi-source causal inference using control variates. *arXiv preprint arXiv:2103.16689*, 2021.

Jinyong Hahn. On the role of the propensity score in efficient semiparametric estimation of average treatment effects. *Econometrica*, pp. 315–331, 1998.

Tobias Hatt, Jeroen Berrevoets, Alicia Curth, Stefan Feuerriegel, and Mihaela van der Schaar. Combining observational and randomized data for estimating heterogeneous treatment effects, 2022.

James J. Heckman, Hidehiko Ichimura, and Petra Todd. Matching as an econometric evaluation estimator. *The review of economic studies*, 65(2):261–294, 1998.

Keisuke Hirano, Guido W. Imbens, and Geert Ridder. Efficient Estimation of Average Treatment Effects Using the Estimated Propensity Score. *Econometrica*, 71(4):1161–1189, 2003.

Julian C. Hong. Strategies to Turn Real-world Data Into Real-world Knowledge. *JAMA Network Open*, 4(10):e2128045, 10 2021.

Guido Imbens, Nathan Kallus, Xiaojie Mao, and Yuhao Wang. Long-term Causal Inference Under Persistent Confounding via Data Combination, 2024. arXiv:2202.07234.

Guido W. Imbens. Nonparametric estimation of average treatment effects under exogeneity: A review. *Review of Economics and statistics*, 86(1):4–29, 2004.

Fredrik D. Johansson, Uri Shalit, and David Sontag. Learning Representations for Counterfactual Inference, 6 2018. arXiv:1605.03661.

E. R. John, K. R. Abrams, C. E. Brightling, and N. A. Sheehan. Assessing causal treatment effect estimation when using large observational datasets. *BMC Medical Research Methodology*, 19(1): 207, 12 2019.

Alistair E.W. Johnson. MIMIC-III, a freely accessible critical care database | Scientific Data, 2016.

Nathan Kallus, Aahlad Manas Puli, and Uri Shalit. Removing Hidden Confounding by Experimental Grounding. *Advances in neural information processing systems*, 2018.

Edward H. Kennedy. Semiparametric theory and empirical processes in causal inference, 7 2016. arXiv:1510.04740.

Edward H. Kennedy. Towards optimal doubly robust estimation of heterogeneous causal effects, 8 2023. arXiv:2004.14497.

Thomas Kneib, Alexander Silbersdorff, and Benjamin Säfken. Rage Against the Mean – A Review of Distributional Regression Approaches. *Econometrics and Statistics*, 26(C):99–123, 2023.

Mark J. van der Laan and Daniel Rubin. Targeted Maximum Likelihood Learning. *The International Journal of Biostatistics*, 2(1), 12 2006.

Lauren D Liao, Emilie Højbjerre-Frandsen, Alan E Hubbard, and Alejandro Schuler. Prognostic adjustment with efficient estimators to unbiasedly leverage historical data in randomized trials. *arXiv preprint arXiv:2305.19180*, 2023.

Jerzy Neyman. Sur les applications de la thar des probabilities aux experiences Agaricales: Essay des principle. Excerpts reprinted (1990) in English. *Statistical Science*, 5(463-472):4, 1923.

Miruna Oprescu, Jacob Dorn, Marah Ghoummaid, Andrew Jesson, Nathan Kallus, and Uri Shalit. B-learner: Quasi-oracle bounds on heterogeneous causal effects under hidden confounding. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 26599–26618. PMLR, 23–29 Jul 2023.

Sky Qiu, Lars van der Laan, and Mark van der Laan. *atmle: Adaptive Targeted Minimum Loss-based Estimation for Augmenting Randomized Controlled Trial with Real-World Data.*, 2024. R package version 0.1.0.

James M. Robins and Andrea Rotnitzky. Semiparametric Efficiency in Multivariate Regression Models with Missing Data. *Journal of the American Statistical Association*, 90(429):122–129, 3 1995.

Donald B. Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology*, 66(5):688, 1974.

Donald B. Rubin. *Matched sampling for causal effects*. Cambridge University Press, 2006.

Maresa Schröder, Dennis Frauen, Jonas Schweisthal, Konstantin Hess, Valentyn Melnychuk, and Stefan Feuerriegel. Conformal Prediction for Causal Effects of Continuous Treatments. In *arXiv.org*, 7 2024.

Alejandro Schuler, David Walsh, Diana Hall, Jon Walsh, Charles Fisher, Critical Path for Alzheimer's Disease, Alzheimer's Disease Neuroimaging Initiative, and Alzheimer's Disease Cooperative Study. Increasing the efficiency of randomized trial estimates via linear adjustment for a prognostic score. *The International Journal of Biostatistics*, 18(2):329–356, 2022.

Jonas Schweisthal, Dennis Frauen, Mihaela van der Schaar, and Stefan Feuerriegel. Meta-Learners for Partially-Identified Treatment Effects Across Multiple Environments, 6 2024. arXiv:2406.02464.

Uri Shalit, Fredrik D Johansson, and David Sontag. Estimating individual treatment effect: generalization bounds and algorithms. *International conference on machine learning*, 2017.

Fangzhou Su, Wenlong Mou, Peng Ding, and Martin J. Wainwright. When is the estimated propensity score better? high-dimensional analysis and bias correction, 2023.

Evelina Tacconelli, Anna Gorska, Elena Carrara, Ruth Joanna Davis, Marc Bonten, Alex W. Friedrich, Corinna Glasner, Herman Goossens, Jan Hasenauer, Josep Maria Haro Abad, José L. Peñalvo, Albert Sanchez-Niubo, Anastassja Sialm, Gabriella Scipione, Gloria Soriano, Yazdan Yazdanpanah, Ellen Vorstenbosch, and Thomas Jaenisch. Challenges of data sharing in European Covid-19 projects: A learning opportunity for advancing pandemic preparedness and response. *The Lancet Regional Health – Europe*, 21, 10 2022.

Mark van der Laan, Benkeser David, and Cai Weixin. Efficient Estimation of Pathwise Differentiable Target Parameters with the Undersmoothed Highly Adaptive Lasso, 2021.

Mark van der Laan, Sky Qiu, and Lars van der Laan. Adaptive-TMLE for the Average Treatment Effect based on Randomized Controlled Trial Augmented with Real-World Data, 5 2024. arXiv:2405.07186.

Stefan Wager. Causal Inference: A Statistical Learning Approach, 2024.

Shirly Wang, Matthew B. A. McDermott, Geeticka Chauhan, Michael C. Hughes, Tristan Naumann, and Marzyeh Ghassemi. MIMIC-Extract: A Data Extraction, Preprocessing, and Representation Pipeline for MIMIC-III. In *Proceedings of the ACM Conference on Health, Inference, and Learning*, pp. 222–235, 4 2020. arXiv:1907.08322.

Christopher Winship and Stephen L. Morgan. The estimation of causal effects from observational data. *Annual Review of Sociology*, 25(1):659–706, 8 1999.

Carlos K. H. Wong, Kristy T. K. Lau, Ivan C. H. Au, Sophelia H. S. Chan, Eric H. Y. Lau, Benjamin J. Cowling, and Gabriel M. Leung. Effectiveness of nirmatrelvir/ritonavir in children and adolescents aged 12–17 years following SARS-CoV-2 Omicron infection: A target trial emulation. *Nature Communications*, 15(1):4917, 6 2024.

Shu Yang and Peng Ding. Combining Multiple Observational Data Sources to Estimate Causal Effects. *Journal of the American Statistical Association*, 115(531):1540–1554, 7 2020.

Tijana Zrnic and Emmanuel J. Candès. Cross-prediction-powered inference. *Proceedings of the National Academy of Sciences*, 121(15):e2322083121, 4 2024.

A ADDITIONAL THEORETICAL PROOFS

In the following, we prove the validity of prediction-powered confidence interval in the non-causal setting, Lemma 4.1 and Lemma 5.1.

A.1 PROOFS OF THE SUPPORTING LEMMAS

Proof of Lemma 5.1 Here we refer the proof to Wager (2024) Chapter 2 about the asymptotic normality of IPW estimator. □

A.2 PROOF OF THEOREM 4.2

We show that $\tau \notin \mathcal{C}_\alpha^{\mathrm{PP}}$ with probability at most α; that is,

$$\limsup_{n,N\to\infty} P\left(\mid \hat{\Delta}_\tau + \hat{\tau}_2 \mid > z_{1-\alpha/2}\sqrt{\frac{\hat{\sigma}^2_\Delta}{n} + \frac{\hat{\sigma}^2_{\tau_2}}{N}}\right) \le \alpha. \tag{9}$$

The central limit theorem implies that

$$\sqrt{n}\left(\hat{\Delta}_\tau - \mathbb{E}[\hat{\Delta}_\tau]\right) \Rightarrow \mathcal{N}\left(0, \sigma^2_\Delta\right); \sqrt{N}\left(\hat{\tau}_2 - \mathbb{E}[\hat{\tau}_2]\right) \Rightarrow \mathcal{N}\left(0, \sigma^2_{\tau_2}\right), \tag{10}$$

where σ^2_Δ is the variance of $\hat{\Delta}_i = \tilde{Y}_{\hat{\eta}}(X_i) - \hat{\tau}_2(X_i)$ and $\sigma^2_{\tau_2}$ is the variance of $\hat{\tau}_2(X_i)$. Therefore, by Slutsky's theorem, we get

$$\begin{aligned}\sqrt{N}\left(\hat{\Delta}_\tau + \hat{\tau}_2 - \mathbb{E}[\hat{\Delta}_\tau + \hat{\tau}_2]\right) =& \sqrt{n}\left(\hat{\Delta}_\tau - \mathbb{E}[\hat{\Delta}_\tau]\right)\sqrt{\frac{N}{n}} + \sqrt{N}\left(\hat{\tau}_2 - \mathbb{E}[\hat{\tau}_2]\right) \\ \Rightarrow& \mathcal{N}\left(0, \frac{1}{p}\sigma^2_\Delta + \sigma^2_{\tau_2}\right).\end{aligned} \tag{11}$$

This in turn implies

$$\limsup_{n,N\to\infty} P\left(\left|\hat{\Delta}_\tau + \hat{\tau}_2 - \mathbb{E}[\hat{\Delta}_\tau + \hat{\tau}_2]\right| > z_{1-\alpha/2}\sqrt{\frac{\hat{\sigma}}{N}}\right) \le \alpha, \tag{12}$$

where $\hat{\sigma}$ is a consistent estimate of the variance $\frac{1}{p}\sigma^2_\Delta + \sigma^2_\tau$. We take $\hat{\sigma} = \frac{N}{n}\sigma^2_\Delta + \sigma^2_\tau$; this estimate is consistent since the two terms are individually consistent estimates of the respective variances. We notice that

$$\mathbb{E}\left[\hat{\Delta}_\tau + \hat{\tau}_2\right] = \mathbb{E}\left[\sum_{i=1}^n \tilde{Y}_{\hat{\eta}}(X_i) - \sum_{i=1}^n \hat{\tau}_2(X_i) + \sum_{j=1}^N \hat{\tau}_2(X_j)\right] = \mathbb{E}\left[\sum_{i=1}^n \tilde{Y}_{\hat{\eta}}(X_i)\right] = \tau, \tag{13}$$

where the last step is that putting together Eq. 12 and Eq. 13 together and apply a union bound, we get

$$\limsup_{n,N\to\infty} P\left(\left|\hat{\Delta}_\tau + \hat{\tau}_2 - \mathbb{E}[\hat{\Delta}_\tau + \hat{\tau}_2]\right| > z_{1-\alpha/2}\sqrt{\frac{\hat{\sigma}^2_\Delta}{n} + \frac{\hat{\sigma}^2_{\tau_2}}{N}}\right) \le \alpha. \tag{14}$$

Therefore, $\limsup_{n,N\to\infty} P\left(\mid \hat{\Delta}_\tau + \hat{\tau}_2 \mid > z_{1-\alpha/2}\sqrt{\frac{\hat{\sigma}^2_\Delta}{n} + \frac{\hat{\sigma}^2_{\tau_2}}{N}}\right) \le \alpha.$ □

A.3 PROOF OF VALIDITY OF PREDICTION-POWERED CI

Formally, we consider estimands of the form $\theta^* = \arg\min_{\theta\in\mathbb{R}^p} \mathbb{E}[l_\theta(X_i, Y_i)]$, and $f(X)$ is the pre-trained machine learning model.

We show that $\tau \notin \mathcal{C}_\alpha^{\mathrm{PP}}$ with probability at most α; that is,

$$\limsup_{n,N\to\infty} P\left(\mid \hat{\Delta}_{\theta^*,j} + \hat{g}^f_{\theta^*,j} \mid > z_{1-\alpha/2}\sqrt{\frac{\hat{\sigma}^2_{\Delta,j}(\theta^*)}{n} + \frac{\hat{\sigma}^2_{g,j}(\theta^*)}{N}}, \forall j \in [p]\right) \le \alpha. \tag{15}$$

For each $j \in [p]$, the central limit theorem implies that

$$\sqrt{n}\left(\hat{\Delta}_{\theta^*,j} - \mathbb{E}[\hat{\Delta}_{\theta^*,j}]\right) \Rightarrow \mathcal{N}\left(0, \sigma^2_{\Delta,j}(\theta^*)\right); \sqrt{N}\left(\hat{g}^f_{\theta^*,j} - \mathbb{E}[\hat{g}^f_{\theta^*,j}]\right) \Rightarrow \mathcal{N}\left(0, \sigma^2_{g,j}(\theta^*)\right), \tag{16}$$

where $\sigma^2_{\Delta,j}(\theta^*)$ is the variance of $\hat{g}^f_{\theta^*,j}(X_i, Y_i) - \hat{g}^f_{\theta^*,j}(X_i, f(X_i))$ and $\sigma^2_{g,j}(\theta^*)$ is the variance of $\hat{g}^f_{\theta^*,j}(X_i, f(X_i))$. Therefore, by Slutsky's theorem, we get

$$\begin{aligned}\sqrt{N}\left(\hat{\Delta}_{\theta^*,j} + \hat{g}^f_{\theta^*,j} - \mathbb{E}[\hat{\Delta}_{\theta^*,j} + \hat{g}^f_{\theta^*,j}]\right) =& \sqrt{n}\left(\hat{\Delta}_{\theta^*,j} - \mathbb{E}[\hat{\Delta}_{\theta^*,j}]\right)\sqrt{\frac{N}{n}} + \sqrt{N}\left(\hat{g}^f_{\theta^*,j} - \mathbb{E}[\hat{g}^f_{\theta^*,j}]\right) \\ \Rightarrow& \mathcal{N}\left(0, \frac{1}{q}\sigma^2_{\Delta,j}(\theta^*) + \sigma^2_{g,j}(\theta^*)\right).\end{aligned} \tag{17}$$

This in turn implies

$$\limsup_{n,N\to\infty} P\left(\left|\hat{\Delta}_{\theta^*,j} + \hat{g}^f_{\theta^*,j} - \mathbb{E}[\hat{\Delta}_{\theta^*,j} + \hat{g}^f_{\theta^*,j}]\right| > z_{1-\alpha/(2p)}\sqrt{\frac{\hat{\sigma}_j}{N}}\right) \leq \alpha/p, \tag{18}$$

where $\hat{\sigma}_j$ is a consistent estimate of the variance $\frac{1}{q}\sigma^2_{\Delta,j}(\theta^*) + \sigma^2_{g,j}(\theta^*)$. We take $\hat{\sigma} = \frac{N}{n}\sigma^2_{\Delta,j}(\theta^*) + \sigma^2_{g,j}(\theta^*)$; this estimate is consistent since the two terms are individually consistent estimates of the respective variances. We notice that

$$\mathbb{E}\left[\hat{\Delta}_{\theta^*,j} + \hat{g}^f_{\theta^*,j}\right] = \mathbb{E}\left[g_{\theta^*}(X_i, Y_i) - g_{\theta^*}(X_i, f(X_i)) + g_{\theta^*}(\tilde{X}_i, f(\tilde{X}_i))\right] = \mathbb{E}\left[g_{\theta^*}(X_i, Y_i)\right] = 0, \tag{19}$$

where the last step is that putting together Eq.18 and Eq.19 together and apply a union bound, we get

$$\begin{aligned}&\limsup_{n,N\to\infty} P\left(\exists j \in [p] : \left|\hat{\Delta}_\tau + \hat{\tau}_2 - \mathbb{E}[\hat{\Delta}_\tau + \hat{\tau}_2]\right| > z_{1-\alpha/(2p)}\sqrt{\frac{\hat{\sigma}^2_\Delta}{n} + \frac{\hat{\sigma}^2_{\tau_2}}{N}}\right) \\ &\leq \sum_{j=1}^{p} \limsup_{n,N\to\infty} P\left(\mid \hat{\Delta}_\tau + \hat{\tau}_2 \mid > z_{1-\alpha/(2p)}\sqrt{\frac{\hat{\sigma}^2_\Delta}{n} + \frac{\hat{\sigma}^2_{\tau_2}}{N}}\right) \\ &= \sum_{j=1}^{p} \limsup_{n,N\to\infty} P\left(\mid \hat{\Delta}_\tau + \hat{\tau}_2 \mid > z_{1-\alpha/(2p)}\sqrt{\frac{\hat{\sigma}^2_\Delta}{n} + \frac{\hat{\sigma}^2_{\tau_2}}{N}}\right) \\ &= \sum_{j=1}^{p} \frac{\alpha}{p} \\ &= \alpha.\end{aligned} \tag{20}$$

□

B ADDITIONAL THEORETICAL RESULTS

In this section, we present more theoretical results about our methods. It can be generalized to deal with distribution shifts and finite sample situations.

B.1 DISTRIBUTION SHIFT

In our main paper, we focus on computing prediction-powered confidence intervals when the $\mathcal{D}^1$ and $\mathcal{D}^1$ come from the same distribution. Herein, we extend out tolls to the case where $\mathcal{D}^1$ comes from $\mathbb{P}$ and the $\mathcal{D}^2$ comes from $\mathbb{Q}$, and these are related by a distribution shift of covariates.

First, we assume that $\mathbb{Q}$ is a known *covariate shift* of $\mathbb{P}$. That is , if we denote by $\mathbb{Q} = \mathbb{Q}_X \cdot \mathbb{Q}_{A|X} \cdot \mathbb{Q}_{Y|A,X}$ and $\mathbb{P} = \mathbb{P}_X \cdot \mathbb{P}_{A|X} \cdot \mathbb{P}_{Y|X}$ the relevant marginal and conditional distributions, we assume that $\mathbb{Q}_{Y|A,X} = \mathbb{P}_{Y|A,X}$, and $\mathbb{Q}_{A|X} = \mathbb{P}_{A|X}$. As in previous sections, we consider the ATE from the

$\mathbb{Q}$ distribution,

$$\hat{\tau} = \mathbb{E}_{\mathbb{Q}}\left[\frac{1}{N}\sum_{j=1}^{N}\hat{\tau}_2(X_j)\right]. \tag{21}$$

The estimand from equation 21 can be related to form on $\mathbb{P}$ using the Radon-Nikodym derivative. In particular, suppose that $\mathbb{Q}_X$ s dominated by $\mathbb{P}_X$ and assume that the Radon-Nikodym derivative $w(X) = \frac{\mathbb{Q}_X}{\mathbb{P}_X}(X)$ is known. Then, we can rewrite equation 21 as,

$$\tau_2^w = \mathbb{E}_{\mathbb{P}}\left[\frac{1}{N}\sum_{j=1}^{N}\hat{\tau}_2(X_j)w(X_j)\right]. \tag{22}$$

In one word, the estimation of $\hat{\tau}$ on $\mathbb{Q}$ can be written as a reweighted function. This permits inference on the rectifier to be based on data sampled from $\mathbb{P}$ as before. For concreteness, we explain the estimation approach in detail. Let,

$$\Delta_\tau^w = \mathbb{E}_{\mathbb{P}}\left[\frac{1}{n}\sum_{j=1}^{n}\tilde{Y}_{\hat{\eta}}(X_i)w(X_i) - \hat{\tau}(X_i)w(X_i)\right]. \tag{23}$$

The confidence interval for the above rectifier suffices for prediction-powered inference on τ.

Confidence interval (Covariate shift). Let $\hat{\sigma}^2_{\tau_2^w}$ denotes empirical variance of $\hat{\tau}_2^w(X)$, $\hat{\sigma}^2_{\Delta^w}$ denotes empirical variance of $\hat{\Delta}_{\tau^w}$. Then, for significance level $\alpha \in (0,1)$, the prediction-powered confidence interval is

$$\mathcal{C}_\alpha^{\mathrm{PP}} = \left(\hat{\tau}^{\mathrm{PP}} \pm z_{1-\frac{\alpha}{2}}\sqrt{\frac{\hat{\sigma}^2_{\Delta^w}}{n} + \frac{\hat{\sigma}^2_{\tau_2^w}}{N}}\right), \tag{24}$$

where $\hat{\tau}^{\mathrm{PP}} = \hat{\Delta}_{\tau^w} + \hat{\tau}_2^w = \frac{1}{n}\sum_{j=1}^{n}\left[\tilde{Y}_{\hat{\eta}}(X_i)w(X_i) - \hat{\tau}(X_i)w(X_i)\right] + \frac{1}{N}\sum_{j=1}^{N}\hat{\tau}_2(X_j)w(X_j)$, $\hat{\sigma}^2_{\Delta^w} = \frac{1}{n}\sum_{i=1}^{n}\left[\tilde{Y}_{\hat{\eta}}(X_i)w(X_i) - \hat{\tau}_2(X_i)w(X_i) - \hat{\Delta}_{\tau^w}\right]^2$, $\hat{\sigma}^2_{\tau_2^w} = \frac{1}{N}\sum_{j=1}^{N}\left[\hat{\tau}_2(X_j)w(X_j) - \hat{\tau}_2^w\right]^2$.

B.2 INFERENCE ON FINITE POPULATION

The method developed in this paper can be directly translated to the *finite-population setting*. Here, we treat $\mathcal{D}^1$ and $\mathcal{D}^2$ as a fixed finite population consisting of n confounder-outcome pairs, without imposing distributional assumptions on the data points. The only assumption required to apply the latter is that $\tilde{Y}_{\hat{\eta}}(X_i) - \hat{\tau}(X_i)$ has a known bound, i.e. $[a_i, b_i]$, valid for all $i \in [n]$.

In the finite-population setting, we still follow the same way of constructing the prediction-powered estimation of ATE,

$$\hat{\tau}^{\mathrm{PP}} = \hat{\Delta}_\tau + \hat{\tau}_2 = \frac{1}{n}\sum_{j=1}^{n}\left[\tilde{Y}_{\hat{\eta}}(X_i) - \hat{\tau}(X_i)\right] + \frac{1}{N}\sum_{j=1}^{N}\hat{\tau}_2(X_j). \tag{25}$$

Confidence interval (Finite population). Let $\hat{\sigma}^2_{\tau_2}$ denotes empirical variance of $\hat{\tau}_2(X)$, $\hat{\sigma}^2_\Delta$ denotes empirical variance of $\hat{\Delta}_\tau$. Then, for significance level $\alpha \in (0,1)$, by Hoeffding's inequality, the prediction-powered confidence interval is

$$\mathcal{C}_\alpha^{\mathrm{PP}} = \left(\hat{\tau}^{\mathrm{PP}} \pm \left[\sqrt{\frac{\sum_{i=1}^{n}(b_i - a_i)^2}{2n^2}\log\frac{2}{\alpha}} + z_{1-\frac{\alpha}{2}}\sqrt{\frac{\hat{\sigma}^2_{\tau_2}}{N}}\right]\right), \tag{26}$$

where $\hat{\sigma}^2_\Delta = \frac{1}{n}\sum_{i=1}^{n}\left(\tilde{Y}_{\hat{\eta}}(X_i) - \hat{\tau}_2(X_i) - \hat{\Delta}_\tau\right)^2$, and $\hat{\sigma}^2_{\tau_2} = \frac{1}{N}\sum_{j=1}^{N}\left(\hat{\tau}_2(X_j) - \hat{\tau}_2\right)^2$.

B.3 INFERENCE ON AVERAGE POTENTIAL OUTCOME

In this section, we show a generalization of our methods to the average potential outcome (APO). We define the mean outcome function in $\mathcal{D}^1$ as $f_1^a(X) := \mathbb{E}[Y(a) \mid X]$.

ATE estimation. Let $\hat{f}_1^a(X)$ be the estimated potential outcome function on $\mathcal{D}^1$ and $\hat{f}_2^a(X)$ be the estimated potential outcome function on $\mathcal{D}^2$. Let rectifier Δ_a denotes difference of $\hat{f}_1^a(X)$ and $\hat{f}_2^a(X)$ on $\mathcal{D}^1$, $\Delta_a = \mathbb{E}\left[\hat{f}_1^a(X) - \hat{f}_2^a(X)\right]$, and $\hat{\Delta}_i = \hat{f}_1^a(X_i) - \hat{f}_2^a(X_i)$. Then, the prediction-powered estimation of APO on $\mathcal{D}^1$ is defined as,

$$\hat{\mu}_{a,1}^{\mathrm{PP}} = \hat{\Delta} + \hat{\mu}_{a,2} = \frac{1}{n}\sum_{i=1}^{n} \hat{\Delta}_i + \frac{1}{N}\sum_{j=1}^{N} \hat{f}_2^a(X_j) = \frac{1}{n}\sum_{i=1}^{n}\left[\hat{f}_1^a(X_i) - \hat{f}_2^a(X_i)\right] + \frac{1}{N}\sum_{j=1}^{N} \hat{f}_2^a(X_j). \tag{27}$$

Confidence interval. Let $\hat{\sigma}_{a,2}^2$ denotes empirical variance of $\hat{f}_2^a(X)$, $\hat{\sigma}_\Delta^2$ denotes empirical variance of $\hat{\Delta}_a$. Then, for significance level $\alpha \in (0, 1)$, the prediction-powered confidence interval is

$$\mathcal{C}_\alpha^{\mathrm{PP}} = \left(\hat{\mu}_{a,1}^{\mathrm{PP}} \pm z_{1-\frac{\alpha}{2}} \sqrt{\frac{\hat{\sigma}_\Delta^2}{n} + \frac{\hat{\sigma}_{a,2}^2}{N}}\right), \tag{28}$$

where $\hat{\sigma}_\Delta^2 = \frac{1}{n}\sum_{i=1}^{n}\left(\hat{f}_1^a(X_i) - \hat{f}_2^a(X_i) - \hat{\Delta}_a\right)^2$, and $\hat{\sigma}_{\tau_2}^2 = \frac{1}{N}\sum_{j=1}^{N}\left(\hat{f}_2^a(X_i) - \hat{\mu}_{a,2}\right)^2$.

C MATHEMATICAL BACKGROUND

We start with a brief overview of PPI (Angelopoulos et al., 2024). In the standard PPI framework, one assumes a labeled dataset $\mathcal{S}_n = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ of n i.i.d. samples drawn from some unknown, but fixed distribution $\mathbb{P}$, where $X_i \in \mathcal{X}$ is input, and $Y_i \in \mathcal{Y}$ is the outome. One further assumes a larger sample $\tilde{\mathcal{S}}_N = \{(\tilde{X}_1, f(\tilde{X}_1)), \ldots, (\tilde{X}_N, f(\tilde{X}_N))\}$ where $n \ll N$, for which the outcome is not available, but where one has access to a pre-trained function $f : \mathcal{X} \to \mathcal{Y}$.

PPI protocol:[8] The objective is then to estimate a statistical quantity of interest given by the estimand θ^* (e.g., the mean). In PPI, one then constructs a prediction-powered estimate $\hat{\theta}^{\mathrm{PP}}$ through a decompisition $\hat{\theta}^{\mathrm{PP}} = m_\theta + \sigma_\Delta$, where m_θ is called 'measure of fit' and Δ_θ is called 'rectifier'. Of note, m_θ is typically defined by the statistical quantity of interest (e.g., m_θ computes the sample average when θ^* is the mean), while the rectifier is a measure of the prediction accuracy of f. Yet, the rectifier is *not* given 'out-of-the-box' but it needs to be carefully derived for the statistical quantity of interest. Finally, the prediction-powered CI is constructed via $\mathcal{C}_\alpha^{\mathrm{PP}} = \{\theta \mid |m_\theta + \Delta_\theta| \leq w_\theta(\alpha)\}$ where $w_\theta(\alpha)$ is a constant that depends on the confidence level. Then $\mathcal{C}_\alpha^{\mathrm{PP}}$ is guaranteed to contain the true parameter θ^* with probability at least $1 - \alpha\%$ (Angelopoulos et al., 2024). Crucially, the prediction-powered CI is smaller than the classical CI when the model f is sufficiently accurate.

Of note, the rectifier must be carefully tailored for the estimand, and the derivation is typically non-trivial, especially in order to obtain further theoretical guarantees (e.g., to show that the CIs are asymptotically valid).

D ALGORITHM

In our main paper, we presented the algorithm for computing the prediction-powered estimation of ATE and confidence interval by combining observational datasets. Below, we state another algorithm 2 that is applicable if $\mathcal{D}^1$ is the RCT dataset. In this case, all kinds of CATE estimation in $\mathcal{D}^1$ could be used.

Algorithm 2 Prediction-powered ATE estimation with combined observational datasets

```
Input: small dataset $\mathcal{D}^1 = (X^1, A^1, Y^1)$, large dataset $\mathcal{D}^2 = (X^2, A^2, Y^2)$, significance level $\alpha \in (0,1)$
$\hat{\tau}_2(X) \leftarrow$ estimate CATE estimator from $\mathcal{D}^2$ and $\tilde{Y}_{\hat{\eta}}(X) \leftarrow$ estimate non-centered influential function score from $\mathcal{D}^1$
$\hat{\Delta}_i \leftarrow \tilde{Y}_{\hat{\eta}}(X_i) - \hat{\tau}_2(X_i)$
$\hat{\tau}_2 \leftarrow \frac{1}{N}\sum_{i=1}^{N} \hat{\tau}_2(X_i)$, and $\hat{\Delta}_\tau \leftarrow \frac{1}{n}\sum_{i=1}^{n} \hat{\Delta}_i$
$\hat{\tau}^{\mathrm{PP}} \leftarrow \hat{\tau}_2 - \hat{\Delta}_\tau$                                      ▷ prediction-powered estimator
$\hat{\sigma}_{\tau_2}^2 \leftarrow \frac{1}{N}\sum_{i=1}^{N} (\hat{\tau}_2(X_i) - \hat{\tau}_2)^2$          ▷ empirical variance of CATE estimation in $\mathcal{D}^2$
$\hat{\sigma}_\Delta^2 \leftarrow \frac{1}{n}\sum_{i=1}^{n} \left(\hat{\Delta}_i - \hat{\Delta}_\tau\right)^2$          ▷ empirical variance of rectifier in $\mathcal{D}^1$
$w_\alpha \leftarrow z_{1-\frac{\alpha}{2}} \sqrt{\frac{\hat{\sigma}_\Delta^2}{n} + \frac{\hat{\sigma}_{\tau_2}^2}{N}}$                          ▷ normal approximation
Output: prediction-powered confidence interval $\mathcal{C}_\alpha^{\mathrm{PP}} = \left(\hat{\tau}^{\mathrm{PP}} \pm w_\alpha\right)$
```

---

[8] We further provide formal derivation in Appendix A.3.

E EXTENDED LITERATURE REVIEW

In this section, we present extended related work on constructing CIs for ATE with multiple datasets in more detail.

The classical way of constructing confidence intervals by making use of one observational dataset is utilizing the TMLE estimator and the AIPW estimator (Bang & Robins, 2005; Laan & Rubin, 2006). It is based on its property of unbiasedness and bounded variance which provide the theoretical support for the valid CIs.

Other works about combining observational datasets to estimate ATEs (Yang & Ding, 2020; Guo et al., 2021). However, they make assumptions that the small dataset needs to be sampled from the main observational dataset. Also, they aim to provide a more efficient estimator of ATE but only provide a bootstrap way to construct confidence intervals which leads to more estimation uncertainty.

Another important stream of combining multiple datasets is that Kallus et al. (2018) proposed a method that combined the RCT dataset and observational dataset to get estimation of CATE, which could be seen as a special case of our method in Sec.5. Although Demirel et al. (2024) also applies prediction-powered inference but focuses on average potential outcomes and does not consider the uncertainty quantification of estimation as we do. We show in Section B.3 that the generalization of our method to the APO.

F EXPERIMENTATION DETAILS

F.1 SYNTHETIC DATASET GENERATION

Following the setup in section 6.1, we consider the three different scenarios of confounding in $\mathcal{D}^2$. As shown in the equation 6.1, a larger value of α_u and a smaller value of l_u implies stronger confounding components. For scenario 1, we set $\alpha_u = 0$ and $l_u = 10^6$ which is the scenario that is almost without confounding. For scenario 2, the $\alpha_u = 0$ and $l_u = 0.5$ which means also we still do not consider the linear component of U but let the unobserved confounder play a more important role in the exponential term. For scenario 3, we set $\alpha_u = 10$ and $l_u = 0.5$ where the unobserved confounder both influences linear and exponential terms, which is the most confounded scenario. For a more clear understanding, we conclude three kernel settings in equation 29,

$$\begin{aligned}
k_{\text{scenario1}}\left(\left(X,U\right),\left(X',U'\right)\right) &= \exp\left[-\frac{(X-X')}{2\times 10^6} - \frac{(U-U')}{2\times 10^6}\right], \\
k_{\text{scenario2}}\left(\left(X,U\right),\left(X',U'\right)\right) &= \exp\left[-\frac{(X-X')}{2\times 10^6} - \frac{(U-U')}{1}\right], \\
k_{\text{scenario3}}\left(\left(X,U\right),\left(X',U'\right)\right) &= 10\times UU' + \exp\left[-\frac{(X-X')}{2\times 10^6} - \frac{(U-U')}{1}\right].
\end{aligned} \tag{29}$$

Here we also need to make clear that the unobserved confounder only plays a role in the generation process of treatment, which means it does not have a straight relationship with the divergence of mean in $\mathcal{D}^2$ and $\mathcal{D}^1$.

F.2 MEDICAL DATASET

We showcase our method on the MIMIC-III dataset (Johnson, 2016), which includes electronic health records (EHRs) from patients admitted to intensive care units. We extract 8 confounders (heart rate, sodium, red blood cell count, glucose, hematocrit, respiratory rate, age, gender) and a binary treatment (mechanical ventilation) using an open-source preprocessing pipeline (Wang et al., 2020). We define the outcome variable as the red blood cell count after treatment. To extract features from the patient trajectories in the EHRs, we sample random time points and average the value of each variable over the ten hours prior to the sampled time point. All samples with missing values and outliers are removed from the dataset. Our final dataset contains 14719 samples, which we still separate dataset by the constant ratio that $n/N = 1/50$ and add noise on $\mathcal{D}^2$.

For the second semi-synthetic dataset, we consider the COVID-19 hospitalizations in Brazil across different regions (Baqui et al., 2020). We are interested in predicting the effect of comorbidity on the mortality of COVID-19 patients. For the environments, we use the regions of the hospitals in Brazil, which are split into North and Central-South. As observed confounders, we include age, sex, and ethnicity. Further, we exclude patients younger than 20 or older than 80 years. To define comorbidity as a binary variable, we define comorbidity as 1 if at least one of the following conditions were diagnosed for the patient: cardiovascular diseases, asthma, diabetes, pulmonary disease, immunosuppression, obesity, liver diseases, neurological disorders, and renal disease. We then use the same data generation process to generate A_i and Y_i, while using the second confounding scenario and keeping the ratio of sample size $n/N = 1/50$.

F.3 IMPLEMENTATION DETAILS

We choose DR-learner as $\hat{\tau}_2$ in $\mathcal{D}^2$ with linear regression and logistic regression model as our basic model. Also, linear regression and logistic regression model for the nuisance function regression when estimating the $\hat{\tau}^{\mathrm{AIPW}}$. We use all default settings for those regression models and we did not perform any hyperparameter optimization, as our method aims to provide an agnostic confidence interval applicable to all CATE estimators. All the experiments are done by five random seeds.

G ADDITIONAL EXPERIMENTAL RESULTS

G.1 NEURAL INSTANTIATIONS OF OUR METHOD

We follow the same experiment setting and data generation process in Section 6.1, but replace the regression model for the nuisance regression model with a multi-layer perception (MLP) in Figure 6. Compared with the simple linear regression, our method achieves CIs that have a shorter width (as desired). Further, the CIs from our method consistently cover the oracle ATE, which again confirms the superiority of our method.

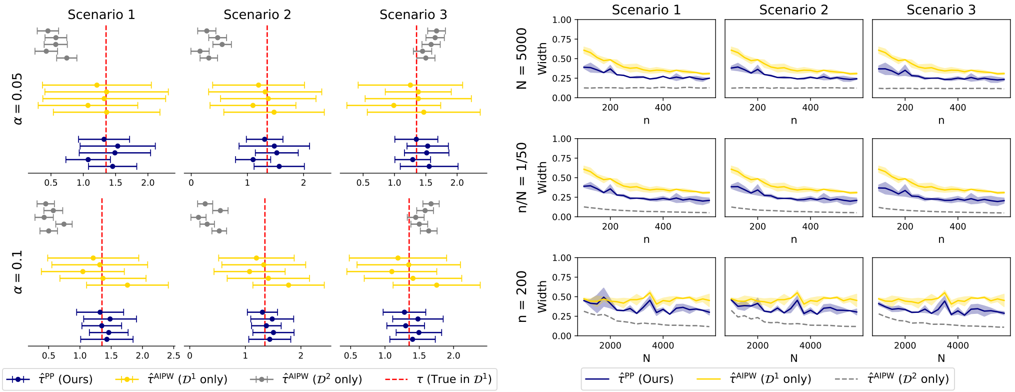


Figure 6: Results for MLP as regression method.

G.2 INSTANTIATION WITH OTHER MACHINE LEARNING MODELS

Figure 7, we follow the same data generated setting as experiments in the Figure 4 but replace the regression method used for the nuisance parameter estimation from a linear regression to XGBoost. Again, our method is highly effective, which demonstrates the flexibility of our method beyond a simple regression model.

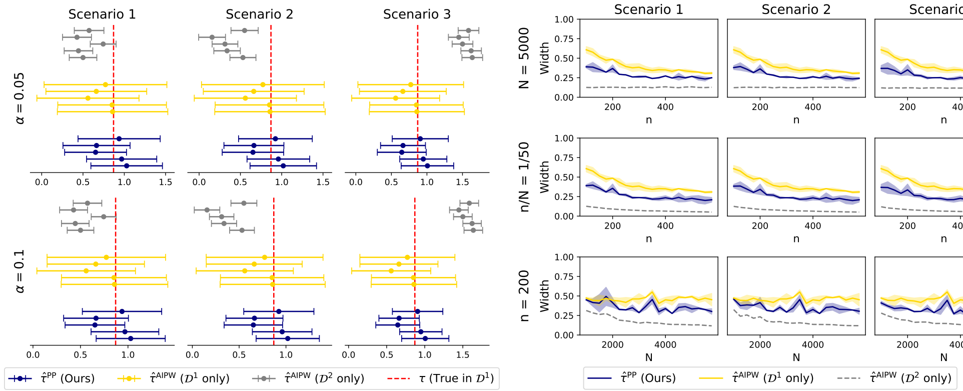


Figure 7: Results for using XGBoost as nuisance parameter regression model.

G.3 HIGH-DIMENSIONAL COVARIATES

We repeated our experiments with more input variables to show that our method is robust in settings with high-dimensional covariate spaces. For this, we used a data-generating mechanism similar to that in the main paper but where we now generate $x \in [-1, 1]^5$ in Figure 8. The results show that the CIs from our method consistently cover the oracle ATE and that our method reduces the width of CIs (as desired).

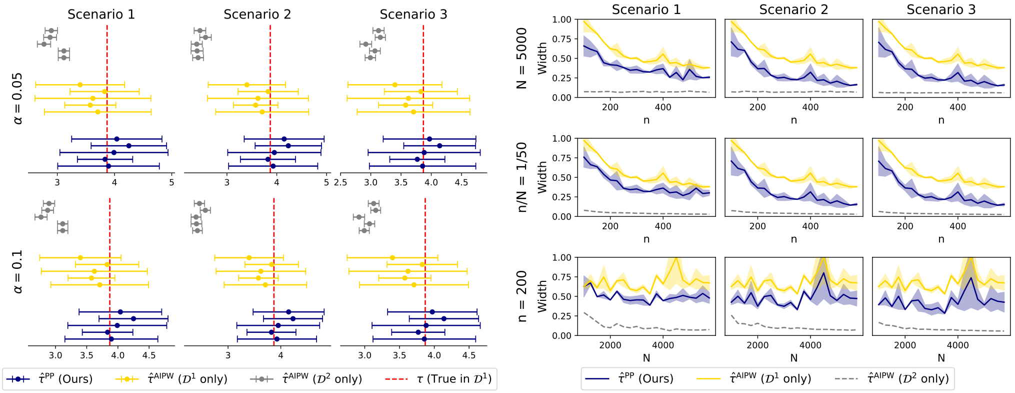


Figure 8: Results for high-dimensional covariates.

G.4 STRENGTH OF DEPENDENCE

As extended experiments based on Section G.3, we simulated the $x \in [-1, 1]^4$ and let $x_5 = \frac{1}{n}\sum_{i=1}^{4} x_i$ which leads to the collinearity in the input space in Figure 9. Thereby, we can assess the sensitivity of our method to a varying strength of dependence in the input space. Compared with i.i.d. high-dimensional covariates, we notice that the dependence does not affect our method. Our method still outperforms the other baselines and achieves the best CI width.

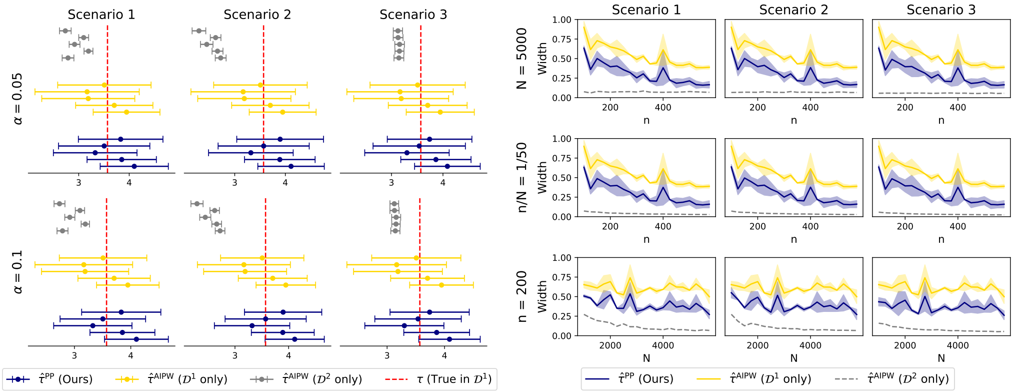


Figure 9: Results for varying dependence strength in input space.

G.5 DIFFERENT STRENGTHS OF (UN)CONFOUNDING IN $\mathcal{D}^1$

We aim to show the experiment setting when relaxing ‘unconfoundedness’ assumption for $\mathcal{D}^1$. We fixed the confoundedness in $\mathcal{D}^2$ as in Scenario 2 but varied the confoundedness in $\mathcal{D}^1$ from Scenario 1 to 3 in Figure 10. We noticed that, while the strength of confounding becomes larger, our method performs better. The results again confirm and that our method performs best.

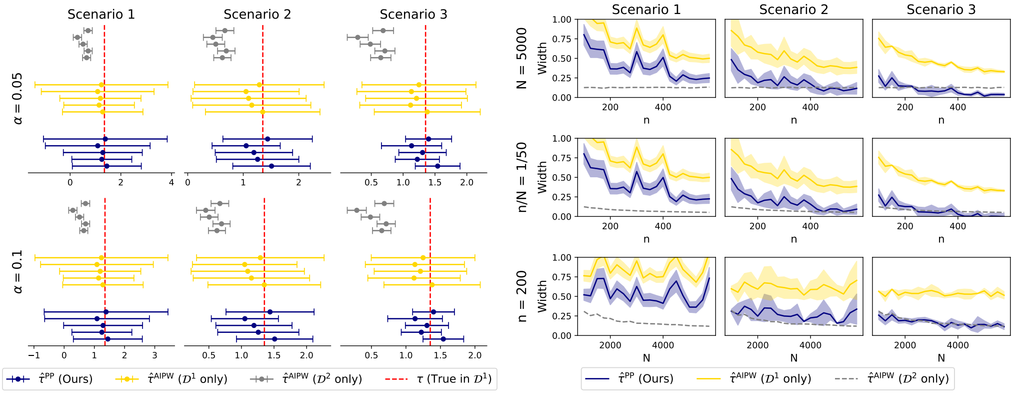


Figure 10: Results for relaxing unconfoundedness assumptions in $\mathcal{D}_1$.

G.6 ROBUSTNESS CHECK OF APPLYING OUR AIPW METHOD TO RCT+OBSERVATIONAL DATASETS

Data: We adopt the same data-generating process as outlined in the main paper while applying our proposed AIPW method described in Section 4 to the RCT+observational setting.

Main results: In Figure 11, we demonstrates that, when replacing the known propensity score with the estimated propensity score, the performance difference is small. Both methods consistently cover the oracle ATE (in the left figure) and show a large gain compared to the naïve baseline (in the right figure).

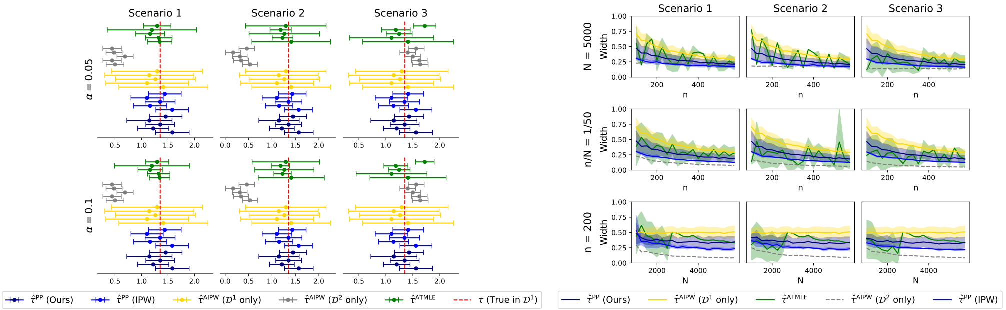


Figure 11: Applying our AIPW method to RCT+observational datasets.

G.7 REFUTATION CHECK OF APPLYING THE A-TMLE METHOD TO OBSERVATIONAL+OBSERVATIONAL DATASETS

We apply the A-TMLE method to the synthetic datasets with observational+observational data. Of note, this violates the assumptions that underly A-TMLE, so we expect that the method leads to large errors.

In Figure 12, we notice that, while applying the A-TMLE method to the synthetic dataset, the A-TMLE performs not that well. Although it constructs the short CIs, it barely covers the oracle ATE in the left figure. In the right figure, the A-TMLE method shows the instability of the estimating process again. These findings highlight that A-TMLE leads to CIs that are *not* faithful in RCT+observational settings. Again, this is expected.

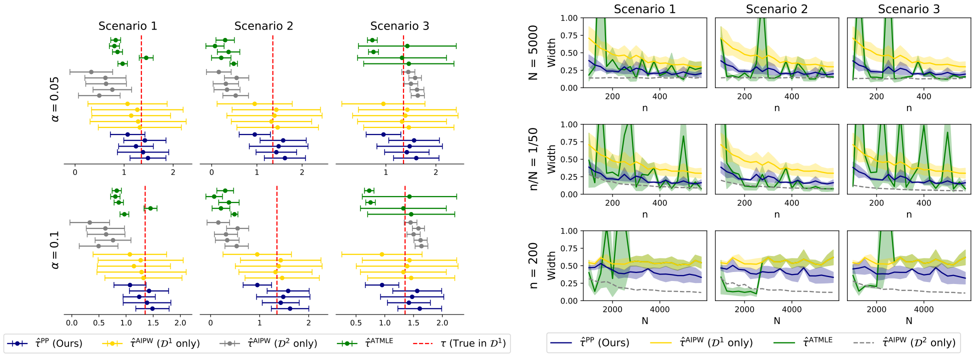


Figure 12: Applying the A-TMLE to multiple observational datasets.

G.8 INCREASING SAMPLE SIZE IN $\mathcal{D}^1$

Data: To provide a more comprehensive evaluation of our method, we increased the sample size in $\mathcal{D}^1$ to enable further comparisons under varying conditions. In Figure 13, the sample size in $\mathcal{D}^2$ is fixed at 5000 ($N = 5000$), while the sample size in $\mathcal{D}^1$ varies from 100 to 2500 across three distinct scenarios. This setup allows us to systematically assess the performance of our method under different data regimes.

Main results: Figure 13 reveals that our method consistently outperforms the naïve method across all scenarios. Notably, as the sample size in $\mathcal{D}^1$ increases, the performance gap gradually narrows, indicating diminishing returns in improvement as more data becomes available in $\mathcal{D}^1$. These results are expected and, therefore, further validate the robustness of our method.

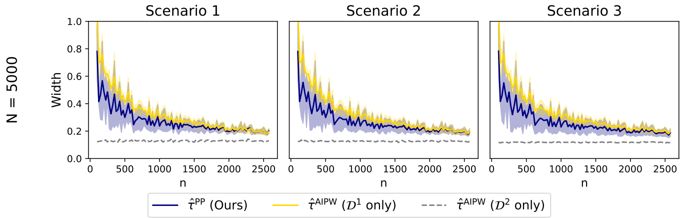


Figure 13: **Performance for an increasing sample size of $\mathcal{D}^1$.** The figure shows the width of the CIs averaged over five different seeds ($\alpha = 0.05$). Here, we vary the size of $\mathcal{D}^1$ datasets given constant sample size N ($\mathcal{D}^2$) from 100 to 2500. Note that $\hat{\tau}^{\mathrm{AIPW}}$ ($\mathcal{D}^2$ only) is shown in intentionally shown in gray: it is *not* faithful as seen in the left plot and therefore *not* a valid baseline. $\Rightarrow$ Our method continually performs better than the $\hat{\tau}^{\mathrm{AIPW}}$ ($\mathcal{D}^2$ only).

G.9 RMSE AND COVERAGE FOR THE EXPERIMENTS WITH SYNTHETIC DATA

In Table 2, we report the RMSE of our point estimation and the width of the CIs in Sec. 6.1.

Table 2: We report the RMSE of the ATE estimator and the width of the CIs. We use the synethic dataset. The results for $\hat{\tau}^{\mathrm{AIPW}}$ ($\mathcal{D}^2$ only) are shown in gray because the estimator is *not* faithful and therefore also *not* a viable baseline. Reported is the average performance over 5 random seeds.

Dataset	RMSE	Width
$\hat{\tau}^{\mathrm{AIPW}}$ ($\mathcal{D}^1$ only)	0.298/0.298/0.298	0.241/0.240/0.237
$\hat{\tau}^{\mathrm{AIPW}}$ ($\mathcal{D}^2$ only)	0.442/0.478/0.476	0.217/0.144/0.131
$\hat{\tau}^{\mathrm{PP}}$ **(Ours)**	**0.276/0.274/0.271**	**0.241/0.240/0.237**

Smaller is better. Best value in bold.

H COMPARISON TO A-TMLE

In this section, we will compare our method against A-TMLE (van der Laan et al., 2024) and thereby highlight key differences as well as why the training in A-TMLE is unstable.

About A-TMLE: A-TMLE is a method that combines two datasets to estimate the ATE and can also construct valid confidence intervals. In van der Laan et al. (2024), the authors prove that the A-TMLE estimator is $\sqrt{n}$-consistent and asymptotically normal and gives valid confidence intervals. With the help of the observational datasets, A-TMLE achieves smaller mean-squared errors and narrower confidence intervals.

The A-TMLE method proceeds as follows. First, in A-TMLE, the author decomposed targeted ATE estimand as the difference of (a) the pooled-ATE estimand $\tilde{\Psi}$ and (b) a bias estimand $\Psi^{\#}$, $\Psi = \tilde{\Psi} - \Psi^{\#}$. At a high level, A-TMLE constructs two separated TMLE estimators for the $\tilde{\Psi}$ and $\Psi^{\#}$. Then, A-TMLE calculates the difference of TMLE outcomes as the targeted estimand.

More specifically, for the bias estimand $\Psi^{\#}$, the estimation process can be decomposed into two steps: (i) learning a parametric working model, and (ii) constructing an efficient estimator for the targeted estimands. In the first step (i), with the 'atmle' R-package (Qiu et al., 2024), the method applies the highly adaptive lasso minimum-loss estimator (HAL-MLE)(van der Laan et al., 2021) with the HAL basis functions for the semi-parametric regression working model. Given the above definition, one can define the working-model-specific projection parameter as

$$\Psi^{\#}(P) = E_p\Pi_p(0 \mid W, 0)\tau_{w,n,\beta(P)}(W, 0) - E_p\Pi_p(0 \mid W, 1)\tau_{w,n,\beta(P)}(W, 1), \tag{30}$$

where P denotes the distribution, W denotes the covariates. We refer to van der Laan et al. (2024) for more details about the notation.

After that, we need the canonical gradient of the $\beta(P)$-component to construct the canonical gradient of the working-model-specific projection parameter $\Psi_{\mathcal{M}_{w,2}}(P)$ at P. However, when calculating the canonical gradient of the $\beta(P)$-component, one of the important things to observe is that $I_p = E_p\Pi(1-\pi)(1 \mid W, A)\phi\phi^T(W, A)$, which measures the variance-covariance structure across basis functions. The expression for I_p adjusts for variability in different directions, reducing weights for directions with high variance (overrepresented in data) and increasing weights where variance is low (underrepresented). Hence, A-TMLE essentially performs an adaptive weighting to make the patients in both datasets more similar for the final estimate.

The reason for why A-TMLE is unstable: The computation of I_p has an important **shortcoming**: When (a) the dimension of the covariate space is low or (b) when collinearity among the covariates exists, the computation of I_p is challenging due to the matrix inversion. Eventually, this can lead to numerical instabilities, which can cause the entire A-TMLE method to break down.

Differences to our method: In addition to the drawbacks of A-TMLE, two key differences exist as follows: (1) differences in ATE estimation processes and (2) differences in the flexibility of estimating $\hat{\tau}_2$. In the following, we discuss the differences (1) and (2) in detail:

(1) *Differences in the ATE estimation process*. One of the key differences between our method and A-TMLE is that our method is based on two different ATE estimations when computing the rectifier. In our method, we define the rectifier Δ_τ as the difference of $\hat{\tau}_{\text{AIPW}}$ and $\hat{\tau}_2$ on $\mathcal{D}^1$. Formally, we have

$$\hat{\Delta}_\tau = \frac{1}{n}\sum_{i=1}^{n}\left[\tilde{Y}_{\hat{\eta}}(x_i) - \hat{\tau}_2(x_i)\right] \tag{31}$$

$$= \frac{1}{n}\sum_{i=1}^{n}\left[\left(\frac{A_i}{\hat{\pi}(x_i)} - \frac{1-A_i}{1-\hat{\pi}(x_i)}\right)Y_i - \frac{A_i - \hat{\pi}(x_i)}{\hat{\pi}(x_i)\left(1-\hat{\pi}(x_i)\right)}\left[\left(1-\hat{\pi}(x_i)\right)\hat{\mu}_1(x_i) + \hat{\pi}(x_i)\hat{\mu}_0(x_i)\right] - \hat{\tau}_2(x_i)\right].$$

In contrast, A-TMLE defines the target estimand by applying a bias correction $\Psi^{\#}$, which can be viewed as the expectation of a weighted combination of the conditional effect of the treatment indicators on the treatment effect of the two treatment arms, where the weights are the probabilities of enrolling in the RCT of the two arms. Then, the highly adaptively lasso minimum-loss estimator (HAL-MLE) is used to learn the semi-parametric regression model.

(2) *Flexibility*. Another key difference is that our method is more flexible, allowing us to use any approach to estimate $\hat{\tau}_2$ in $\mathcal{D}^2$. In contrast, the process in A-TMLE is more rigid: A-TMLE constructs a TMLE for the pooled-ATE and bias correction term estimation. This can limit the flexibility for computing $\hat{\tau}_2$, especially when we want to use different modeling approaches for both datasets (which is likely given that one dataset is probably larger than the other!).

Instead, our method supports a variety of approaches, allowing end-users of our method to better adapt to the underlying data-generating process. For example, we can use various meta-learners like the S-learner, T-learner, R-learner, and DR-learner, where each comes with unique strengths in practice. The S-learner, for instance, works well when there are fewer treatment interactions, while the T-learner and R-learner handle more complex treatment effect patterns.

Additionally, our method allows us to use pre-trained models directly (which is unlike A-TMLE!). This allows us – in our method – to calculate the ATE from model predictions without needing to re-fit or modify the model. Alternatively, one can even use large language models or foundation models to generate the predictions of $\hat{\tau}_2$. The flexibility to use various models or integrate pre-trained models makes our approach more flexible to handle a broad variety different settings and data structures. We believe that this makes our method a powerful tool for accurate ATE estimation in a range of applications. For example, if we are given a pre-trained machine learning model $f(x)$, then we have access to the predictions on $\mathcal{D}^2$ as $\hat{f}(x)$. Formally, we then yield the measure of fit and the rectifier via

$$\hat{\tau}_2 = \frac{1}{N}\sum_{j=1}^{N}\hat{f}(x_j), \tag{32}$$

$$\begin{aligned}\hat{\Delta}_\tau = & \frac{1}{n}\sum_{i=1}^{n}\left[\tilde{Y}_{\hat{\eta}}(x_i) - \hat{f}(x_i)\right] = \frac{1}{n}\sum_{i=1}^{n}\Bigg[\left(\frac{A_i}{\hat{\pi}(x_i)} - \frac{1-A_i}{1-\hat{\pi}(x_i)}\right)Y_i \\ & - \frac{A_i - \hat{\pi}(x_i)}{\hat{\pi}(x_i)\left(1-\hat{\pi}(x_i)\right)}\left[\left(1-\hat{\pi}(x_i)\right)\hat{\mu}_1(x_i) + \hat{\pi}(x_i)\hat{\mu}_0(x_i)\right] - \hat{f}(x_j)\Bigg],\end{aligned} \tag{33}$$

$$\hat{\tau}^{\mathrm{PP}} = \frac{1}{N}\sum_{j=1}^{N}\hat{f}(x_j) + \frac{1}{n}\sum_{i=1}^{n}\left[\tilde{Y}_{\hat{\eta}}(x_i) - \hat{f}(x_i)\right]. \tag{34}$$

According to the central limited theorem of the predictions $f(x)$ and the asymptotical normality of the AIPW estimator, we can construct valid CI as we mentioned in the main paper. This means, we have $\mathcal{C}_\alpha^{\mathrm{PP}} = \left(\hat{\tau}^{\mathrm{PP}} \pm z_{1-\frac{\alpha}{2}}\sqrt{\frac{\hat{\sigma}_\Delta^2}{n} + \frac{\hat{\sigma}_{\tau_2}^2}{N}}\right)$, where $\hat{\sigma}_\Delta^2$ and $\hat{\sigma}_{\tau_2}^2$ are variance of the rectifier and measure of fit respectively.