PRIVATE: Differentially Private Confidence Intervals for Average Treatment Effects

Maresa Schröder
LMU Munich
Munich Center for Machine Learning
maresa.schroeder@lmu.de

Justin Hartenstein
University of Mannheim
jhartens@mail.uni-mannheim.de

Stefan Feuerriegel
LMU Munich
Munich Center for Machine Learning
feuerriegel@lmu.de

Abstract

The average treatment effect (ATE) is widely used to evaluate the effectiveness of drugs and other medical interventions. In safety-critical applications like medicine, reliable inferences about the ATE typically require valid uncertainty quantification, such as through confidence intervals (CIs). However, estimating treatment effects in these settings often involves sensitive data that must be kept private. In this work, we present PRIVATE, a novel machine learning framework for computing CIs for the ATE under differential privacy. Specifically, we focus on deriving *valid privacy-preserving CIs for the ATE* from observational data. Our PRIVATE framework consists of three steps: (i) estimating a differentially private ATE through output perturbation; (ii) estimating the differentially private variance through a truncated output perturbation mechanism; and (iii) constructing the CIs while accounting for the uncertainty from both the estimation and privatization steps. Our PRIVATE framework is model agnostic, doubly robust, and ensures valid CIs. We demonstrate the effectiveness of our framework using synthetic and real-world medical datasets. To the best of our knowledge, we are the first to derive a general, doubly robust framework for valid CIs of the ATE under (ε, δ)-differential privacy.

1 Introduction

Estimating the *average treatment effect (ATE)* from observational data is highly relevant for evaluating the effectiveness of drugs and other medical interventions [e.g., 7, 13, 25]. To ensure the reliability of ATE estimates, one often needs to "move beyond the mean" by accounting for the estimation uncertainty [31, 43]. This is typically achieved through *confidence intervals (CIs)*, which provide a range within which the true effect is likely to lie [49]. Medical studies frequently report 95% CIs to show the precision of the estimated treatment effects. For example, the clinical trial for the Moderna COVID-19 vaccine specified that the effectiveness must meet a predefined threshold based on the 95% CI, and, ultimately, the trial reported a 95% CI of 89.3% to 96.8% [6].

Estimating ATEs in medicine typically involves sensitive patient data [12], that must be kept private. Many regulations have been introduced to enforce data privacy, such as the US Health Insurance Portability and Accountability Act (HIPAA) and the EU General Data Protection Regulation (GDPR). One approach to analyse sensitive data is to use privacy mechanisms such as *differential privacy (DP)*, enabling population-level inference while protecting individual information [20]. Due to strong theoretical guarantees, especially (ε, δ)-DP has received increasing attention [e.g., 1, 10].

In this work, we aim to *compute differentially private CIs for the ATE*. However, this is *non-trivial*: constructing CIs solely based on the private ATE estimate does *not* suffice to ensure DP of the entire CI, as estimating the variance required for constructing CIs involves additional queries to the underlying data. Thus, it is necessary to develop privacy-preserving mechanisms not only for the ATE estimate itself but also for the variance estimate that is used to construct the CIs. In other words, *both components must be computed in a differentially private manner* to ensure privacy.

Yet, this introduces an additional *challenge*: *privatizing the ATE and its variance introduces additional noise*. This added uncertainty needs to be accounted for when computing CIs to ensure *validity*[1]. Failure to do so results in CIs that do not maintain their stated coverage probability. More formally, in estimation under DP, uncertainty arises from two sources: (i) the sampling error inherent to the data and the uncertainty introduced by the machine learning (ML) model, and (ii) the noise added to ensure privacy. The second source of uncertainty makes its quantification for DP estimates *inherently different from standard uncertainty quantification*. Therefore, traditional methods to construct CIs fail to account for the latter, resulting in CIs that <u>not</u> valid and thus give rise to misleading conclusions.

So far, there are only few works on treatment effect estimation under DP [e.g., 29, 44, 52]. Existing works mainly focus on estimating causal quantities as point estimates, whereas privacy-preserving CIs are missing. As discussed above, extending existing methods for ATE estimation under DP to construct differentially private CIs is <u>not</u> readily feasible. Other studies focus on privacy-preserving CIs [e.g., 9, 24], yet outside of treatment effect estimation and, therefore, cannot be directly applied to causal quantities such as the ATE. Guha and Reiter [29] and Ohnishi and Awan [52] present private CIs for the ATE, but in *different* settings (e.g., by making restrictive assumptions on the data-generating process, by being limited to linear models, and/or by focusing on a different definition of DP). To our knowledge, a method for computing valid CIs for the ATE under (ε, δ)-DP is missing.

Here, we develop PRIVATE, the first framework designed to compute *valid* CIs for the ATE under (ε, δ)-DP (see Figure 1). Our framework is built upon the augmented inverse probability weighting (AIPW) estimator, which offers several practical advantages such as being model-agnostic, doubly robust, and Neyman-orthogonal [54, 64]. Of note, model-agnostic means that our framework can be used together with *arbitrary* machine learning models such as neural networks. To obtain the CIs, we further introduce a novel variance estimator for the ATE that is based on truncated output perturbation and satisfies DP. This allows us to construct *valid*, model-agnostic CIs even in the presence of privatization noise.
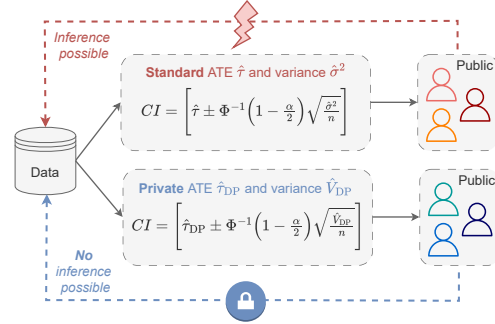

Figure 1: **Standard vs. private CIs for the ATE.** Standard CIs allow for inference about individual data samples, whereas CIs under DP are designed to prevent such inference.

Our main contributions:[2] (1) We propose PRIVATE, a novel ML framework to construct differentially private CIs for the ATE. (2) We prove that our framework guarantees DP and provides valid CIs. (3) We perform extensive experiments using synthetic and real-world medical data to demonstrate the effectiveness and validity of our framework.

2 Related work

We briefly summarize the existing literature relevant to our work, namely (i) ATE estimation under DP, (ii) differentially private CI construction, and (iii) CIs for the ATE. We give an overview in Fig. 2.

Differentially private ATE estimation: Research on privacy mechanisms for ATE estimation is scarce. Lee et al. [44] design a privacy-preserving Horvitz-Thompson type estimator by privatizing the propensity score function. However, due to the employed regularization, the privatized ATE estimate is eventually biased, which prevents the construction of valid CIs. Guha and Reiter [29] provide a DP method for the estimation of ATEs of binary outcomes. Other works have analyzed

---

[1] A CI is valid if it achieves its nominal coverage rate asymptotically. For example, a 95% confidence interval is valid if it covers the true parameter approximately 95% of the time under repeated sampling.

[2] We provide all code and data in our GitHub repository `https://github.com/m-schroder/PrivATE`

private treatment effect estimation in other DP regimes like local DP [2, 51] or label DP [37], or focus on other causal quantities [e.g., 52, 60, 67]. However, as explained in our introduction, computing differentially private CIs for the ATE is non-trivial.

Differentially private CIs for non-causal quantities: Several works focus on the construction of differentially private CIs, yet *outside* of causal inference. Here, one stream constructs CIs for quantities estimated through parametric models where the underlying population is known and/or normally distributed [e.g., 40, 19, 24]. Other works have focused on constructing CIs for the parameters of a trained model, e.g., linear regression [9, 61]. A more general method has been proposed to construct asymptotic confidence intervals for parameters of any model trained by empirical risk minimization [65]; however, this method imposes strong assumptions on the convexity and the smoothness of the loss function, making it incompatible with many machine learning models. In sum, the above works focus on non-causal quantities and, therefore, are <u>not</u> directly applicable to ATE estimation. Therefore, we must first derive an asymptotically normal differentially private ATE estimate before we can then obtain private CIs through our framework.

CIs for the ATE: Several works construct CIs for the ATE based on $\sqrt{n}$-consistent estimators with asymptotic coverage guarantees [8, 34]. Some works also focus on finite-sample CIs. However, these require significantly stronger assumptions on the data [4]. Finally, [66] uses prediction-powered inference to derive CIs for the ATE from multiple datasets. However, <u>no</u> method ensures DP.

	References	Model agnostic*	Neural models	Data agnostic*	(ε, δ)-DP	Doubly robust / Neyman orthogonal
WATE	Guha & Reiter (2024)	✓	✓	Restrictive data generating process	X	X
ATE	Ohnishi & Awan (2024)	X	X	✓	X	X
ATE	**Ours**	✓	✓	✓	✓	✓

* Model agnostic methods can be applied to any ML model. Data agnostic implies no restrictions on the data structure (e.g., binary outcomes).

Figure 2: Key literature for estimating CIs of the ATE.

Literature on *privacy mechanisms* for CIs of the ATE is scarce. We are aware of only two works [29, 52], yet both works have *clear limitations*. Guha and Reiter [29] focus on the WATE (i.e., the weighted ATE, which, in principle, can be transformed into the ATE), but the method puts restrictive assumptions on the data-generating process, because of which the method is later <u>not</u> applicable. [52] is limited to a sieve logistic regression and is <u>not</u> designed to deal with neural networks.

Importantly, both works focus on a *different* privacy definition, namely, pure ε-DP, while we focus on the more common (ε, δ)-DP. On top of that, unlike our PRIVATE framework, <u>none</u> of the two methods benefit from the favorable properties of being doubly robust and Neyman-orthogonal.

Research gap: To the best of our knowledge, we are the first to provide a general machine learning framework to construct valid CIs for the ATE under (ε, δ)-DP.

3 Problem formulation

Notation: We denote random variables by capital letters X and their realizations by small letters x. We refer to the probability distribution over X by P_X, where we omit the subscript whenever it is apparent from the context. The probability mass function for discrete X is given by $P(x) = P(X = x)$ and the probability density function w.r.t. the Lebesgue measure by $p(x)$.

Setting: We consider a dataset $D := \{(X_i, A_i, Y_i)\}_{i=1,\dots,n}$, consisting of observed confounders X, a binary treatment $A \in \{0, 1\}$, and a bounded outcome $Y \in \mathcal{Y}$, where $Z_i := (X_i, A_i, Y_i) \sim P$ i.i.d., $Z_i \in \mathcal{Z}$, and $\mathcal{X}, \mathcal{Y}$ have bounded domains. We use the potential outcomes framework [57] and denote the potential outcome of intervention a by $Y(a)$. Furthermore, we define the propensity score as $\pi(x) := P(A = 1 \mid X = x)$ and the outcome function as $\mu(x, a) := \mathbb{E}[Y \mid X = x, A = a]$.

Objective: We estimate the average treatment effect (ATE) τ under (ε, δ)-DP and construct CIs for $\tau := \mathbb{E}[Y(1) - Y(0)]$ so that the CIs (i) capture both estimation and privatization uncertainty and (ii) retain the DP guarantees. We make the standard assumptions in causal inference: positivity, consistency, and unconfoundedness [e.g., 17, 57]. Then, the ATE is identified as $\tau = \mathbb{E}[Y \mid A = 1] - \mathbb{E}[Y \mid A = 0]$. We provide an introduction to ATE estimation and an overview of the corresponding estimators in Supplement A.

In the following, we derive differentially private CIs for the ATE estimated via the AIPW estimator. The AIPW estimator has several advantages [55, 64]: it is unbiased and asymptotically normally distributed. Furthermore, the AIPW estimator is *doubly robust* and Neyman-orthogonal, making it

insensitive to small errors in the estimated nuisances $\hat{\mu}$ and $\hat{\pi}$. However, as we show in the following, it is *non-trivial* to construct private CIs from the AIPW estimator.

3.1 Differential privacy

To ensure the privacy of sensitive information, (ε, δ)-differential privacy (DP) guarantees that the inclusion or exclusion of any individual data point does not affect the estimated outcome by more than a pre-specified threshold defined by the *privacy budget ε and δ* [20, 21]. Specifically, the probability density of any outcome y on dataset $D \in \mathcal{Z}^n$ must be ε-*indistinguishable* from the probability density of the same outcome y on a neighboring dataset $D' \in \mathcal{Z}^n$ with a probability of at least $1 - \delta$.

Definition 3.1 (Differential privacy [21]). Two datasets D and D' are called *neighbors*, denoted as $D \sim D'$, if their Hamming distance equals one, i.e., $d_{\mathrm{H}}(D, D') = 1$. A function $\mathbf{f}_D : \mathcal{X}^d \mapsto \mathbb{R}^d$ trained on dataset D is (ε, δ)-*differentially private* if, for all neighboring datasets $D, D' \in \mathcal{Z}^n$, all measurable $S \subseteq \mathbb{R}^d$ and all $\mathbf{x} \in \mathcal{X}^d$, it holds that

$$P(\mathbf{f}_D(\mathbf{x}) \in S) \leq \exp(\varepsilon) \cdot P(\mathbf{f}_{D'}(\mathbf{x}) \in S) + \delta. \tag{1}$$

DP can be achieved by employing the so-called *Gaussian privacy mechanisms* [22], where the aim is to perturb the prediction by adding appropriately calibrated zero-centered Gaussian noise so that two predictions resulting from neighboring databases cannot be differentiated [e.g., 15, 68].

Definition 3.2 (Gaussian privacy mechanism [22]). Let $\mathbf{f} : \mathcal{X}^d \mapsto \mathbb{R}^d$ and let $\Delta_2(\mathbf{f}) = \sup_{D \sim D', \mathbf{x} \in \mathcal{X}^d} ||\mathbf{f}_D(\mathbf{x}) - \mathbf{f}_{D'}(\mathbf{x})||$ denote the l_2-sensitivity if $\mathbf{f}$. Let $\mathbf{U} \sim \mathcal{N}(0, \sigma^2)$ for $\sigma \geq \frac{1}{\varepsilon}\sqrt{2 \ln(1.25/\delta)} \cdot \Delta_2(\mathbf{f})$. Then, $\mathbf{f}_D^{\mathrm{DP}}(\mathbf{x}) = \mathbf{f}_D(\mathbf{x}) + \mathbf{U}$ is (ε, δ)-differentially private.

3.2 Problem statement

Uncertainty quantification for the ATE is commonly presented in terms of confidence intervals.[3] In our work, we aim at *differentially private uncertainty quantification* of the ATE. Specifically, we aim to derive a (ε, δ)-differentially private estimate $\hat{\tau}_{\mathrm{DP}}$ and the corresponding differentially private CI

$$\mathrm{CI}_{\mathrm{DP}} := \left[\hat{\tau}_{\mathrm{DP}} \pm \Phi^{-1}\left(1 - \frac{\alpha}{2}\right)\sqrt{\frac{\hat{V}_{\mathrm{DP}}}{n}}\right] \tag{2}$$

for a confidence level $(1 - \alpha)$, where $\hat{V}_{\mathrm{DP}}$ denotes the *augmented differentially private variance estimate* capturing the variation stemming from *both* the estimation and the privatization process. Importantly, estimating $\hat{V}_{\mathrm{DP}}$ is *non-trivial*: setting $\hat{V}_{\mathrm{DP}} = \hat{\sigma}_{\mathrm{DP}}^2$ for a privatized variance estimate of the ATE would lead to *undercoverage* of the final private CI (see Section 4). Therefore, we must carefully derive the augmented variance estimate $\hat{V}_{\mathrm{DP}}$ to guarantee valid CIs.

4 Our PRIVATE framework

4.1 Overview

Our framework for constructing (ε, δ)-differentially private CIs for the ATE proceeds in three steps: ① **ATE privatization $\hat{\tau}_{\mathrm{DP}}$:** We estimate the ATE $\hat{\tau}$ through doubly robust estimation and guarantee DP via output perturbation ($\rightarrow$ our Theorem 4.3). ② **Differentially private variance estimation $\hat{\sigma}_{\mathrm{DP}}$:** To be able to construct valid CIs under DP, we additionally obtain an estimate of the variance of $\hat{\tau}$ satisfying DP guarantees ($\rightarrow$ our Theorem 4.5). ③ **Constructing CIs for the ATE under DP:** We finally employ
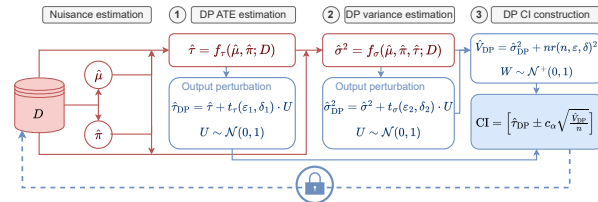

Figure 3: **Our proposed PRIVATE framework** for constructing (ε, δ)-differentially private CIs for the ATE.

---

[3]For an asymptotically consistent ATE estimate $\hat{\tau}$ and its estimated variance $\hat{\sigma}^2$, the CI for a confidence level $(1 - \alpha), \alpha \in (0, 1)$, is given by $\mathrm{CI} = \left[\hat{\tau} \pm \Phi^{-1}\left(1 - \frac{\alpha}{2}\right)\sqrt{\frac{\hat{\sigma}^2}{n}}\right]$, where $\Phi^{-1}\left(1 - \frac{\alpha}{2}\right)$ denotes the critical value given by the respective quantile of the standard normal distribution.

4

$\hat{\tau}_{\mathrm{DP}}$ and $\hat{\sigma}_{\mathrm{DP}}$ to construct differentially private CIs for the ATE. Notably, the resulting CIs capture the uncertainty stemming from the non-private estimation in Step ① as well as the additional uncertainty arising due to the privatization in Step ② ($\rightarrow$ our Theorem 4.6). We show our framework in Figure 3.

Of note, our PRIVATE framework is highly flexible. We do not make any assumptions about the parametric form of the model or the data distribution. Our framework thus is *model-agnostic* and can be seamlessly instantiated with any machine learning model. Furthermore, our framework applies to observational and experimental data as we do *not* assume the propensity function to be known.

4.2 Mathematical preliminaries

We focus on the AIPW estimator in our derivations as it (i) consistently estimates the ATE and (ii) is asymptotically normal, allowing us to derive valid CIs. The AIPW estimator is given by

$$\hat{\tau} = \frac{1}{n} \sum_{i=1}^{n} \left[\hat{\mu}(X, 1) - \hat{\mu}(X, 0) + \frac{Y_i - \hat{\mu}(X, 1)}{\hat{\pi}(X_i)} A_i - \frac{Y_i - \hat{\mu}(X, 0)}{1 - \hat{\pi}(X_i)} (1 - A_i) \right]. \tag{3}$$

Our work exploits a connection between model robustness and privacy in terms of the *influence function* of the underlying estimator following former work in this field [e.g., 21, 60].

Definition 4.1. Let T be a functional of a distribution with $T(P)$ the parameter of interest. The *influence function* (IF) of T at data point z under distribution P is defined as

$$\mathrm{IF}(z, T; P) := \lim_{t \mapsto 0} \frac{T((1-t)P + t\delta_z) - T(P)}{t}, \tag{4}$$

where δ_z denotes the Dirac-delta functional at z. The *gross-error sensitivity* of T is defined as

$$\gamma(T, P) := \sup_{z \in \mathcal{Z}} \|\mathrm{IF}(z, T; P)\|. \tag{5}$$

Lemma 4.2. *The IF of the AIPW learner is given by the influence function of the final mean estimation in the second step. Specifically, the influence stemming from training the nuisance estimators π and μ can be neglected during the privatization step if the estimators are sufficiently stable.*[4]

Proof. We prove Lemma 4.2 in Supplement B. $\qquad\square$

Lemma 4.2 is of great benefit: Due to the construction of the mean estimation, the final AIPW estimator is *not* sensitive to small perturbations in the nuisance functions. This allows us to employ the complete data for the second-stage averaging, as we do not need to privatize the nuisance estimates. Together with its consistency and asymptotic normality, the AIPW estimator is a perfect underlying method for providing differentially private confidence intervals of the ATE.

Remark: To avoid stability guarantees, one can alternatively privatize the nuisances by a DP-method of choice. The privatization neither affects the efficiency of the estimator [60] nor the validity of the CIs. Specifically, we split the dataset D into two disjoint subsets D_1 and D_2. On D_1, we estimate $\hat{\pi}, \hat{\mu}$ under $(\varepsilon_1/2, \delta_1/2)$-DP, which are then employed to estimate $\hat{\tau}$ on D_2. Note that data splitting is common to prevent overfitting and guarantee efficiency of the estimators [e.g., 26, 64].

4.3 Framework

① Differentially private ATE estimation

In the first step, we privatize the original ATE estimate $\hat{\tau}$. Directly employing Definition 3.2 requires calculating the model sensitivity, which can be computationally expensive or even infeasible. We instead exploit the fact that a suitably scaled gross-error sensitivity γ can upper bound the global sensitivity [5]. Therefore, we first derive the influence function of the AIPW estimator so that we are then able to privatize $\hat{\tau}$ based on γ.

---

[4] We require the base estimators for μ and π to be sufficiently stable in terms of uniform stability [11]. This is necessary to provide fast convergence of the (possibly misspecified) nuisance estimators. It is worth noting that many common ML methods fulfill this criterion.

For the AIPW estimator, the *influence function score* is given by [54]

$$\Gamma_{\hat{\eta}}(z) = \left(\frac{a_i}{\hat{\pi}(x_i)} - \frac{1-a_i}{1-\hat{\pi}(x_i)} \right) y_i - \frac{(1-\hat{\pi}(x_i))\,\hat{\mu}_1(x_i) + \hat{\pi}(x_i)\hat{\mu}_0(x_i)}{\hat{\pi}(x)\,(1-\hat{\pi}(x_i))} (a_i - \hat{\pi}(x_i)).$$

Since the AIPW estimator can equivalently be written as $\hat{\tau} = \sum_{i=1}^{n} \Gamma_{\hat{\eta}}(z_i)$, the IF of the AIPW estimator directly follows via[5]

$$\mathrm{IF}^{\mathrm{AIPW}}(z, \tau_0; P) = \Gamma_{\hat{\eta}}(z) - \hat{\tau}. \tag{6}$$

Theorem 4.3 (ATE privatization)**.** *Let $z := (a, x, y)$ define a data sample following the joint distribution $\mathcal{Z}$, and let $\hat{\eta} = (\hat{\mu}, \hat{\pi})$ be the estimated nuisance functions. Furthermore, let D be the training dataset with $|D| = n$. For $U \sim \mathcal{N}(0,1)$, (ε, δ)-DP is fulfilled by*

$$\hat{\tau}_{\mathrm{DP}} := \hat{\tau} + \sup_{z \in \mathcal{Z}} \left\| \Gamma_{\hat{\eta}}(z) - \hat{\tau} \right\| \cdot \frac{5\sqrt{2\ln(n)\ln(2/\delta)}}{\varepsilon n} \cdot U, \tag{7}$$

Proof. We prove Theorem 4.3 in Supplement B. □

② Differentially private variance estimation

Solely constructing CIs based on the private ATE estimate is *not* sufficient to guarantee DP, as estimating the variance necessary for constructing the CIs requires additional queries to the underlying data. Therefore, we must now derive a differentially private estimate of the variance, $\hat{\sigma}^2_{\mathrm{DP}}$, for $\hat{\tau}_{\mathrm{DP}}$.

The variance estimate depends on the estimated nuisance functions. Standard variance estimators are, in general, *only* consistent if *both* nuisance functions are correctly specified [28]. Thus, straightforward estimators, such as the influence function-based estimator [18], are not doubly robust and, therefore, can easily lead to *undercoverage* of the CIs in practice, which is undesirable in safety-critical applications. As a remedy, we estimate $\hat{\sigma}^2_{\mathrm{AIPW}}$ by a *doubly robust variance estimator*.

It is common to obtain such an estimator through nonparametric bootstrapping [27, 62]. However, bootstrapping requires many queries to the data, which is disadvantageous for estimation under DP guarantees, as each query requires a separate privacy budget.

Instead, in our work, we estimate $\hat{\sigma}^2_{\mathrm{AIPW}}$ via the *empirical sandwich estimator* [36]. This has the following benefits: (i) the empirical sandwich estimator is doubly robust, and (ii) it consistently estimates the variance of the AIPW estimator [62]. Importantly, for the AIPW estimator, the sandwich estimator can be directly represented by the IF (see Supplement A)

$$\hat{\sigma}^2_{\mathrm{AIPW}} = \frac{1}{n} \sum_{i=1}^{n} \mathrm{IF}^{\mathrm{AIPW}}(z, \tau_0; P)^2. \tag{8}$$

To derive a differentially private version of $\hat{\sigma}^2_{\mathrm{AIPW}}$, we now follow the same strategy as in Step ①. We first derive the IF of $\hat{\sigma}^2_{\mathrm{AIPW}}$ and then perform output perturbation employing the gross-error sensitivity to upper-bound the smooth sensitivity of $\hat{\sigma}^2_{\mathrm{AIPW}}$.

Lemma 4.4. *The influence function of $\hat{\sigma}^2_{\mathrm{AIPW}}$ is given by*

$$\mathrm{IF}^{\sigma}(z, \sigma_0, P) = (\Gamma_{\hat{\eta}}(z) - \hat{\tau})^2 - \hat{\sigma}^2_{\mathrm{AIPW}}. \tag{9}$$

Proof. We prove Theorem 4.4 in Supplement B. □

Additionally, we observe that the differentially private variance is naturally positive. To impose this restriction during the privatization process, we employ a *truncated privacy mechanism*: Let $\mathbf{f} : \mathcal{X}^d \mapsto \mathbb{R}^d$ and let $\Delta_2(\mathbf{f}) = \sup_{D \sim D', \mathbf{x} \in \mathcal{X}^d} \|\mathbf{f}_D(\mathbf{x}) - \mathbf{f}_{D'}(\mathbf{x})\|$ denote the l_2-sensitivity of $\mathbf{f}$. Furthermore, let $\mathbf{U} \sim \mathcal{N}(0, \sigma^2)$ for $\sigma \geq \frac{1}{\varepsilon}\sqrt{2\ln(1.25/\delta)} \cdot \Delta_2(\mathbf{f})$. Then,

$$\mathbf{f}_D^{\mathrm{DP}}(\mathbf{x}) = \max\{0, \mathbf{f}_D(\mathbf{x}) + \mathbf{U}\} \tag{10}$$

is (ε, δ)-differentially private. We prove the statement in the proof of Theorem 4.5.

---

[5]For an introduction to the construction of IFs for common statistics such as the mean, we refer to [33].

Theorem 4.5 (Differentially private variance estimation). *Let $z := (a, x, y)$ define a data sample following the joint distribution $\mathcal{Z}$, and let $\hat{\eta} = (\hat{\mu}, \hat{\pi})$ be the estimated nuisance functions. Furthermore, let D be the training dataset with $|D| = n$. For $U \sim \mathcal{N}(0, 1)$, we define*

$$\hat{\sigma}_{\mathrm{DP}}^2 := \max\left\{ 0, \hat{\sigma}_{\mathrm{AIPW}}^2 + \sup_{z \in \mathcal{Z}} \left\| \mathrm{IF}^{\sigma}(z, \sigma_0, P) \right\| \cdot \frac{5\sqrt{2\ln(n)\ln(2/\delta)}}{\varepsilon n} \cdot U \right\}. \tag{11}$$

Then, $\hat{\sigma}_{\mathrm{DP}}^2$ is (ε, δ)-differentially private.

Proof. We prove Theorem 4.5 in Supplement B. $\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad$ □

③ Constructing CIs for the ATE under DP

One might think that it is sufficient to simply privatize the variance to ensure DP. However, that is *not* the case: In the following, we clarify that achieving DP for the variance alone would indeed be sufficient—but *only* in the *asymptotic regime*, due to asymptotic normality. However, this leads to problematic implications in finite samples. Specifically, the privatized variance *may underestimate the true variance*, leading to undercoverage in the CIs.

Observe that the AIPW estimator $\hat{\tau}$ is a consistent estimator for the true ATE τ_0 and asymptotically normally distributed with variance σ_{AIPW}^2 [e.g., 64], i.e.,

$$\sqrt{n}\left(\hat{\tau} - \tau_0\right) \to \mathcal{N}\left(0, \sigma_{\mathrm{AIPW}}^2\right). \tag{12}$$

For $\hat{\tau}_{\mathrm{DP}}$, by Slutsky's theorem, the same asymptotic guarantee holds. This implies that the privatization asymptotically does not have an effect, since, for infinite data, no privatization is needed.

In finite settings, however, the privatization introduces additional noise on $\hat{\tau}$ (and thus variance), which we account for in the following when constructing our CIs. This is in line with prior literature postulating that DP inference should be conservative to ensure reliability even in small sample sizes [e.g., 39]. Observe that $(\hat{\tau}_{\mathrm{DP}} - \hat{\tau}) =: B$ is itself a random variable (for fixed $\hat{\tau}$) following distribution $\mathcal{N}\left(0, r(\varepsilon, \delta, n)^2\right)$, where

$$r(\varepsilon, \delta, n) := \sup_{z \in \mathcal{Z}} \left\| \Gamma_{\hat{\eta}}(z) - \hat{\tau} \right\| \cdot \frac{5\sqrt{2\ln(n)\ln(2/\delta)}}{\varepsilon n} \tag{13}$$

and therefore

$$\sqrt{n}\left(\hat{\tau}_{\mathrm{DP}} - \tau_0\right) = \sqrt{n}B + \sqrt{n}\left(\hat{\tau} - \tau_0\right). \tag{14}$$

To incorporate the *finite sample variance* stemming from the privatization, we see from (14) that, for the variance V, it must hold that: $V = \sigma_{\mathrm{AIPW}}^2 + nr(\varepsilon, \delta, n)^2$. ropose a *truncated Gaussian noise privacy mechanism*. For ease of notation, we define the truncated Gaussian distribution as $\mathcal{N}^+(\mu, \sigma)$ with density

To construct the CIs, we further make use of two important properties of DP: the (i) *post-processing* and the (ii) *sequential decomposition* property (see Supplement A for details). The post-processing property states that differentially private estimates cannot be de-privatized by further processing of the estimate. The sequential decomposition property states that the privacy budgets ε_i of multiple sequential estimation tasks i on the same data add up when releasing a joint outcome. For our task, the property (i) justifies that constructing the intervals from differentially private ATE and variance estimates does not require further privatization. Further, property (ii) establishes that we must split the overall privacy budget into two smaller budgets, ε_1 and ε_2, attributed to ATE and variance estimation.

With the properties above and Steps ① and ②, we now state the main contribution of our work.

Theorem 4.6 (Valid CIs for the ATE under DP). *Let $z := (a, x, y)$ define a data sample following the joint distribution $\mathcal{Z}$, and let $\hat{\eta} = (\hat{\mu}, \hat{\pi})$ be the estimated nuisance functions. We define the privacy budget (ε, δ) and the sub-budgets $\varepsilon_1, \varepsilon_2, \delta_1, \delta_2$, for ATE and variance estimation under DP such that $\varepsilon_1 + \varepsilon_2 = \varepsilon$ and $\delta_1 + \delta_2 = \delta$. Furthermore, let D be the training dataset with $|D| = n$, and let $\hat{\tau}_{\mathrm{DP}}$ denote the $(\varepsilon_1, \delta_1)$-differentially private ATE estimate and $\hat{\sigma}_{\mathrm{DP}}$ denote the $(\varepsilon_2, \delta_2)$-differentially private variance estimate. Then, the valid and (ε, δ)-differentially private CI is given by*

$$\mathrm{CI}_{\mathrm{DP}} := \left[\hat{\tau}^{\mathrm{DP}} \pm \Phi^{-1}\left(1 - \frac{\alpha}{2}\right) \sqrt{\frac{\hat{V}_{\mathrm{DP}}}{n}} \right], \tag{15}$$

where $\Phi^{-1}(1 - \alpha/2)$ denotes the respective quantile of the standard normal distribution, γ_τ is the *gross-error sensitivity of the ATE estimation, and*

$$\hat{V}_{\text{DP}} = \hat{\sigma}_{\text{DP}}^2 + \gamma_\tau^2 \cdot \frac{50\ln(n)}{n\varepsilon_1^2} \ln\left(\frac{2}{\delta_1}\right). \tag{16}$$

Proof. We prove Theorem 4.6 in Supplement B. □

5 Experiments

Baselines: There are <u>no</u> other methods for constructing CIs for the ATE under (ε, δ)-DP. Therefore, we cannot compare our PRIVATE against existing baselines, but instead, evaluate our framework in terms of empirical coverage (see below). We further report standard non-private CIs and CIs computed via a *non-private* variance $\hat{\sigma}_{\text{AIPW}}^2$ instead of $\hat{V}_{\text{DP}}$ together with $\hat{\tau}_{\text{DP}}$. We refer to this as naïve CIs in the following. We expect similar coverage for the non-private and our intervals, but undercoverage for the naïve CIs. However, the standard and naïve CIs are *not* private.

Implementation: We evaluate two versions of PRIVATE with different base learners for nuisance estimation: (i) logistic regression for propensity estimation and kernel ridge regression for the outcome function (**Kernel**),(ii) neural networks with $\tanh$ regularization for both nuisances (**NN**) optimized through stochastic gradient descent.

	Base learner	Conf. $(1 - \alpha)$	Coverage		
			Standard CIs	Naïve CIs	PRIVATE (ours)
1	Kernel	0.80	0.792 ± 0.018	0.008 ± 0.004	0.784 ± 0.018
	NN	0.80	0.762 ± 0.019	0.024 ± 0.007	0.824 ± 0.017
	Kernel	0.90	0.902 ± 0.013	0.036 ± 0.008	0.912 ± 0.013
	NN	0.90	0.868 ± 0.015	0.028 ± 0.007	0.894 ± 0.014
	Kernel	0.95	0.960 ± 0.009	0.038 ± 0.009	0.964 ± 0.008
	NN	0.95	0.950 ± 0.010	0.020 ± 0.006	0.962 ± 0.009
2	Kernel	0.80	0.770 ± 0.019	0.004 ± 0.001	0.848 ± 0.016
	NN	0.80	0.800 ± 0.018	0.012 ± 0.005	0.832 ± 0.017
	Kernel	0.90	0.904 ± 0.013	0.008 ± 0.004	0.910 ± 0.013
	NN	0.90	0.912 ± 0.013	0.008 ± 0.004	0.904 ± 0.013
	Kernel	0.95	0.950 ± 0.010	0.010 ± 0.004	0.958 ± 0.009
	NN	0.95	0.952 ± 0.010	0.016 ± 0.006	0.954 ± 0.009

Table 1: **Empirical coverage.** Comparison of PRIVATE against standard and naïve CIs based on Kernel and NN as base learners over 500 runs for $\varepsilon = 0.5$ and $\delta = 10^{-5}$. We report the mean and standard deviation across the runs. To retain as much of the prediction power of $\hat{\tau}^{\text{DP}}$ as possible, we set $\varepsilon_1 = 0.9\varepsilon$, $\delta_1 = 0.9\delta$. In green: valid CIs overlap with the desired confidence level $(1 - \alpha)$ and are thus faithful. $\Rightarrow$ *As expected, we observe similar coverage for the standard CIs and* PRIVATE. *The naïve CIs fail to fulfill any coverage guarantees, underlining the need for accounting for the finite sample variance from the privatization of $\hat{\tau}$.*

Performance metrics: We evaluate PRIVATE based on the notion of *faithfulness* [e.g., 59], i.e., we compute the *empirical coverage* of the CIs and assess whether it surpasses the $1 - \alpha$ threshold.

5.1 Evaluation on synthetic datasets

Evaluating methods in causal ML is difficult, as counterfactual outcomes are never observed in real-world data. Therefore, it is standard to evaluate the methods on synthetic data [e.g., 17]. We use two datasets adapted from Oprescu et al. [53]: • **Dataset 1** consists of two covariates, both of which influence treatment assignment and outcomes. • **Dataset 2** is a more complex dataset consisting of 25 covariates, of which a subset acts as confounders. For details, see Supplement C.

Results: We show that our PRIVATE gives CIs that are *faithful*. The results for both datasets are reported in Table 1 for confidence levels $(1 - \alpha) \in \{0.8, 0.9, 0.95\}$. We also report the coverage of the naïve CIs with *non-private* variance and the standard non-private CIs. Recall that our aim is not to benchmark the methods but rather to show that our framework yields faithful CIs. $\Rightarrow$ We observe that PRIVATE achieves the desired coverage guarantees, which confirms that the CIs are faithful.

Performance across varying privacy budgets: We now vary the privacy budget ε in Figure 4. We observe that PRIVATE behaves as expected: for increasing privacy budgets, the intervals become tighter, eventually converging to the non-private estimate. We note that privacy inevitably induces a trade-off regarding the width of the interval, which we discuss in Supplement D.

Performance across varying sample sizes: We further analyze the performance of PRIVATE across different sample sizes n (see Fig. 4). The results are as expected and confirm our theoretical insights: For larger sample sizes, the CIs become tighter, converging to the asymptotic solution. Additionally,

we observe that PRIVATE is not sensitive to the choice of the base learner: both versions with (i) kernel ridge regression (=Kernel) and (ii) neural network (=NN) lead to similar results. This again demonstrates the flexibility of our framework.

5.2 Evaluation on medical datasets

We demonstrate the real-world applicability of PRIVATE in a medical context on the Right Heart Catheterization (RHC) dataset [16]. The dataset contains observations of 5735 patients admitted to intensive care units. We aim to predict the ATE of receiving a Swan-Ganz catheterization on the probability of death following hospitalization. Further details on the RHC dataset are in Supplement C.

Results: In Table 2, we present the CIs returned by PRIVATE compared to an adapted version of the naïve CIs constructed with a *valid variance*, which includes the correction term for the finite-sample variance stemming from ATE privatization. Again, we do not aim to benchmark the methods as the naïve CIs are non-private. Rather, we aim to understand whether the CIs from PRIVATE are faithful. We observe that the PRIVATE CIs include

Figure 4: **Performance across different privacy budgets and sample sizes.** Results for Dataset 1 over 10 runs with base values $\varepsilon = 0.5$, $\delta = 10^{-5}$, $n = 3000$. The standard CIs only incorporate the sampling uncertainty without privacy considerations. $\Rightarrow$ *The results confirm our theoretical intuition: with a larger budget or a large n, our CIs approach the standard CIs.*

the valid, naïve CIs, which we see as an indication of validity. In other words, our PRIVATE gives CIs that, in addition to fulfilling (ε, δ)-DP, remain informative.

Medical interpretation: We observe that we always receive a positive private treatment effect $\hat{\tau}_{DP}$, indicating an average increase in the probability of death after receiving a Swan-Ganz catheter. The negative effect on survival is likely to be attributed to the life-threatening complications and risks of introducing the catheter, such as pulmonary artery rupture [e.g., 41]. The positive $\hat{\tau}_{DP}$ also coincides with the non-private ATE estimate $\hat{\tau} = 1.17$ and is in line with findings in the literature [e.g., 16, 58].

Conf.	Budget	DP-ATE	Confidence intervals	
$1 - \alpha$	ε	$\hat{\tau}_{DP}$	Valid naïve CIs	PRIVATE (Ours)
	0.10	0.2398	$[-1.1619, 1.6416]$	$[-1.1620, 1.6417]$
0.80	0.25	0.0796	$[-0.4871, 0.6462]$	$[-0.4871, 0.6463]$
	0.50	0.0541	$[-0.2361, 0.3442]$	$[-0.2361, 0.3442]$
	0.10	0.2397	$[-1.5597, 2.0391]$	$[-1.5599, 2.0393]$
0.90	0.25	0.1059	$[-0.5885, 0.8003]$	$[-0.5887, 0.8005]$
	0.50	0.0534	$[-0.3143, 0.4211]$	$[-0.3143, 0.4210]$
	0.10	0.2517	$[-1.8680, 2.3715]$	$[-1.8686, 2.3719]$
0.95	0.25	0.0965	$[-0.7888, 0.9819]$	$[-0.7891, 0.9822]$
	0.50	0.0468	$[-0.3856, 0.4792]$	$[-0.3858, 0.4793]$

Table 2: **Medical application.** Reported are (i) the standard *non-private* CIs and (ii) our PRIVATE CIs on the RHC dataset averaged over 5 runs of the kernel base learner. $\Rightarrow$ *As desired, the* PRIVATE *CIs mainly include the valid but non-private CIs, which demonstrates that* PRIVATE *achieves the desired empirical coverage, suggesting valid CIs.*

Extension to interventional RCT data: Our PRIVATE framework applies to both observational and experimental data. In experimental randomized controlled trials (RCTs), the propensity score is known and thus does not need to be estimated. Note that in some cases, however, estimating the propensity score, even in RCTs, can be beneficial [63]. So far, we have evaluated PRIVATE on observational data. To show the applicability of PRIVATE to RCTs, we conduct additional experiments in Supplement E.

6 Discussion

Impact & limitations: Our work has a direct impact on many practical applications, especially in safety-critical settings. Here, uncertainty quantification is necessary to make reliable and safe inferences. Yet, many applications are subject to strict privacy laws. We propose PRIVATE to
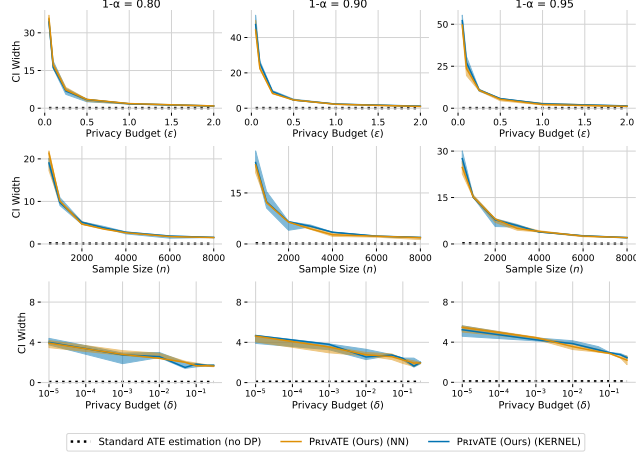
9

construct CIs for the ATE under DP. However, we recommend cautious use in practice as both ATE estimation and privacy mechanisms rely upon mathematical assumptions that need to be fulfilled.

Conclusion: We proposed a new, general framework for constructing differentially private CIs for the ATE. Our PRIVATE framework is carefully designed to capture the additional uncertainty from the privatization and, therefore, return valid CIs. Our PRIVATE framework is flexible: one can use PRIVATE with any machine learning model for learning the propensity and the outcome function.

Acknowledgements

References

[1] M. Abadi, A. Chu, I. Goodfellow, H. B. McMahan, I. Mironov, K. Talwar, and L. Zhang. Deep learning with differential privacy. In *Conference on Computer and Communications Security*, 2016.

[2] A. Agarwal and R. Singh. Causal inference with corrupted data: Measurement error, missing values, discretization, and differential privacy. *arXiv preprint*, 2107.02780, 2024.

[3] A. Alaa and M. van der Schaar. Bayesian inference of individualized treatment effects using multi-task Gaussian processes. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2017.

[4] P. M. Aronow, J. M. Robins, T. Saarinen, F. Sävje, and J. Sekhon. Nonparametric identification is not enough, but randomized controlled trials are. In *Workshop on the Neglected Assumptions in Causal Inference (NACI)*, 2021.

[5] M. Avella-Medina. Privacy-preserving parametric inference: A case for robust statistics. *Journal of the American Statistical Association*, 116(534):969–983, 2021.

[6] L. R. Baden, H. M. El Sahly, B. Essink, K. Kotloff, S. Frey, R. Novak, D. Diemert, S. A. Spector, N. Rouphael, C. B. Creech, J. McGettigan, S. Khetan, N. Segall, J. Solis, A. Brosz, C. Fierro, H. Schwartz, K. Neuzil, L. Corey, P. Gilbert, H. Janes, D. Follmann, M. Marovich, J. Mascola, L. Polakowski, J. Ledgerwood, B. S. Graham, H. Bennett, R. Pajon, C. Knightly, B. Leav, W. Deng, H. Zhou, S. Han, M. Ivarsson, J. Miller, and T. Zaks. Efficacy and safety of the mRNA-1273 SARS-CoV-2 vaccine. *The New England Journal of Medicine*, 384(5): 403–416, 2021.

[7] K. V. Ballmann. Biomarker: Predictive or prognostic? *Journal of Clinical Oncology*, 33(33): 3968–3971, 2015.

[8] H. Bang and J. M. Robins. Doubly robust estimation in missing data and causal inference models. *Biometrics*, 61(4):962–973, 2005.

[9] A. F. Barrientos, J. P. Reiter, A. Machanavajjhala, and Y. Chen. Differentially private significance tests for regression coefficients. *Journal of Computational and Graphical Statistics*, 28(2): 440–453, 2019.

[10] R. Bassily, A. Smith, and A. Thakurta. Private empirical risk minimization: Efficient algorithms and tight error bounds. In *IEEE Annual Symposium on Foundations of Computer Science*, 2014.

[11] O. Bousquet and A. Elisseeff. Stability and generalization. *Journal of Machine Learning Research*, 2:499–526, 2002.

[12] K. B. Brothers and M. A. Rothstein. Ethical, legal and social implications of incorporating personalized medicine into healthcare. *Personalized Medicine*, 12(1):43–51, 2015.

[13] K. G. Buell, A. B. Spicer, J. D. Casey, K. P. Seitz, E. T. Qian, E. J. Graham Linck, W. H. Self, T. W. Rice, P. Sinha, P. J. Young, M. W. Semler, and M. M. Churpek. Individualized treatment effects of oxygen targets in mechanically ventilated critically ill adults. *JAMA*, 331 (14):1195–1204, 2024.

[14] R. H. Byrd, P. Lu, J. Nocedal, and C. Zhu. A limited-memory algorithm for bound-constrained optimization. *SIAM Journal on Scientific Computing*, 16(5):1190–1208, 1995.

[15] K. Chaudhuri, C. Monteleoni, and A. D. Sarwate. Differentially private empirical risk minimization. *Journal of Machine Learning Research*, 12:1069–1109, 2011.

[16] A. F. Connors, T. Speroff, N. V. Dawson, C. Thomas, F. E. Harrell Jr, D. Wagner, N. Desbiens, L. Goldman, A. W. Wu, R. M. Califf, W. J. Fulkerson Jr, H. Vidaillet, S. Broste, P. Bellamy, J. Lynn, and W. A. Knaus. The effectiveness of right heart catheterization in the initial care of critically Ill patients. *JAMA*, 276(11):889–897, 1996.

[17] A. Curth and M. van der Schaar. Nonparametric estimation of heterogeneous treatment effects: From theory to learning algorithms. In *Conference on Artificial Intelligence and Statistics (AISTATS)*, 2021.

[18] J.-C. Deville. Variance estimation for complex statistics and estimators: Linearization and residual techniques. *Survey Methodology*, 25(2):193–203, 1999.

[19] V. D'Orazio and J. Honaker. Differential privacy for social science inference. *Sloan Foundation Economics Research Paper*, 2015.

[20] C. Dwork. Differential privacy. In *International Colloquium on Automata, Languages, and Programming*, 2006.

[21] C. Dwork and J. Lei. Differential privacy and robust statistics. In *ACM Symposium on Theory of Computing*, 2009.

[22] C. Dwork and A. Roth. The algorithmic foundations of differential privacy. *Foundations and Trends in Theoretical Computer Science*, 9(3-4):211–407, 2014.

[23] C. Dwork, F. McSherry, K. Nissim, and A. Smith. Calibrating noise to sensitivity in private data analysis. *Journal of Privacy and Confidentiality*, 7(3):17–51, 2016.

[24] C. Ferrando, S. Wang, and D. Sheldon. Parametric bootstrap for differentially private confidence intervals. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2022.

[25] S. Feuerriegel, D. Frauen, V. Melnychuk, J. Schweisthal, K. Hess, A. Curth, S. Bauer, N. Kilbertus, I. S. Kohane, and M. van der Schaar. Causal machine learning for predicting treatment outcomes. *Nature Medicine*, 30(4):958–968, 2024.

[26] D. J. Foster and V. Syrgkanis. Orthogonal statistical learning. *arXiv preprint*, 1901.09036, 2019.

[27] M. J. Funk, D. Westreich, C. Wiesen, T. Stürmer, M. A. Brookhart, and M. Davidian. Doubly robust estimation of causal effects. *American Journal of Epidemiology*, 173(7):761–767, 2011.

[28] S. Gruber and M. J. van der Laan. tmle: An R package for targeted maximum likelihood estimation. *Journal of Statistical Software*, 51(13):1–35, 2012.

[29] S. Guha and J. P. Reiter. Differentially private estimation of weighted average treatment effects for binary outcomes. *arXiv preprint*, 2408.14766, 2024.

[30] D. Haziza and F. Picard. Doubly robust point and variance estimation in the presence of imputed survey data. *The Canadian Journal of Statistics*, 40(2):259–281, 2012.

[31] J. J. Heckman, J. Smith, and N. Clements. Making the most out of programme evaluations and social experiments: Accounting for heterogeneity in programme impacts. *Review of Economic Studies*, 64(4):487–535, 1997.

[32] K. Hess, V. Melnychuk, D. Frauen, and S. Feuerriegel. Bayesian neural controlled differential equations for treatment effect estimation. In *International Conference on Learning Representations (ICLR)*, 2024.

[33] O. Hines, O. Dukes, K. Diaz-Ordaz, and S. Vansteelandt. Demystifying statistical learning based on efficient influence functions. *The American Statistician*, 76(3):292–304, 2022.

[34] K. Hirano, G. W. Imbens, and G. Ridder. Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica*, 71(4):1161–1189, 2003.

[35] S. Horii and Y. Chikahara. Uncertainty quantification in heterogeneous treatment effect estimation with gaussian-process-based partially linear model. In *Conference on Artificial Intelligence (AAAI)*, 2024.

[36] P. J. Huber. The behavior of maximum likelihood estimates under nonstandard conditions. In *Berkeley Symposium on Mathematical Statistics and Probability*, 1967.

[37] A. Javanmard, V. Mirrokni, and J. Pouget-Abadie. Causal inference with differentially private (clustered) outcomes. *arXiv preprint*, 2308.00957, 2024.

[38] A. Jesson, S. Mindermann, U. Shalit, and Y. Gal. Identifying causal-effect inference failure with uncertainty-aware models. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2020.

[39] V. Karwa and S. Vadhan. Finite sample differentially private confidence intervals. *arXiv preprint*, 1711.03908, 2017.

[40] V. Karwa and S. Vadhan. Finite sample differentially private confidence intervals. In *Innovations in Theoretical Computer Science Conference*, 2018.

[41] T. J. Kearney and M. M. Shabot. Pulmonary artery rupture associated with the swan-ganz catheter. *Chest*, 108(5):1349–1352, 1995.

[42] J. k. Kim and D. Haziza. Doubly robust inference with missing data in survey sampling. *Statistica Sinica*, 2014.

[43] T. Kneib, A. Silbersdorff, and B. Säfken. Rage against the mean: A review of distributional regression approaches. *Econometrics and Statistics*, 26(C):99–123, 2023.

[44] S. K. Lee, L. Gresele, M. Park, and K. Muandet. Privacy-preserving causal inference via inverse probability weighting. *arXiv preprint*, 1905.12592, 2019.

[45] L. Lei and E. J. Candès. Conformal inference of counterfactuals and individual treatment effects. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 83(5):911–938, 2021.

[46] Y. Ma, V. Melnychuk, J. Schweisthal, and S. Feuerriegel. Diffpo: A causal diffusion model for learning distributions of potential outcomes. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2024.

[47] V. Melnychuk, D. Frauen, and S. Feuerriegel. Normalizing flows for interventional density estimation. In *International Conference on Machine Learning (ICML)*, 2023.

[48] V. Melnychuk, S. Feuerriegel, and M. van der Schaar. Quantifying aleatoric uncertainty of the treatment effect: A novel orthogonal learner. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2024.

[49] J. Neyman. Outline of a theory of statistical estimation based on the classical theory of probability. *Philosophical Transactions of the Royal Society of London. Series A, Mathematical and Physical Sciences*, 236:333–380, 1937.

[50] K. Nissim, S. Raskhodnikova, and A. Smith. Smooth sensitivity and sampling in private data analysis. In *ACM Symposium on Theory of Computing*, 2007.

[51] Y. Ohnishi and J. Awan. Locally private causal inference for randomized experiments. *arXiv preprint*, 2301.01616, 2023.

[52] Y. Ohnishi and J. Awan. Differentially private covariate balancing causal inference. *arXiv preprint*, 2410.14789, 2024.

[53] M. Oprescu, V. Syrgkanis, and Z. S. Wu. Orthogonal random forest for causal inference. In *International Conference on Machine Learning (ICML)*, 2019.

[54] J. Robins and A. Rotnitzky. Semiparametric efficiency in multivariate regression models with missing data. *Journal of the American Statistical Association*, 90(429):122–129, 1995.

[55] J. Robins, A. Rotnitzky, and L. P. Zhao. Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association*, 89(427): 846–866, 1994.

[56] P. R. Rosenbaum and D. B. Rubin. Assessing sensitivity to an unobserved binary covariate in an observational study with binary outcome. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, pages 212–218, 1983.

[57] D. B. Rubin. Causal inference using potential outcomes: Design, modeling, decisions. *Journal of the American Statistical Association*, 100(469):322–331, 2005.

[58] J. D. Sandham, R. D. Hull, R. F. Brant, L. Knox, G. F. Pineo, C. J. Doig, D. P. Laporta, S. Viner, L. Passerini, H. Devitt, A. Kirby, and M. Jacka. A randomized, controlled trial of the use of pulmonary-artery catheters in high-risk surgical patients. *The New England journal of medicine*, 348(1):5–14, 2003.

[59] M. Schröder, D. Frauen, J. Schweisthal, K. Heß, V. Melnychuk, and S. Feuerriegel. Conformal prediction for causal effects of continuous treatments. *arXiv preprint*, 2407.03094, 2024.

[60] M. Schröder, V. Melnychuk, and S. Feuerriegel. Differentially private learners for heterogeneous treatment effects. In *International Conference on Learning Representations (ICLR)*, 2025.

[61] O. Sheffet. Differentially private ordinary least squares. In *International Conference on Machine Learning (ICML)*, 2017.

[62] B. E. Shook-Sa, P. N. Zivich, C. Lee, K. Xue, R. K. Ross, J. K. Edwards, J. S. A. Stringer, and S. R. Cole. Double robust variance estimation with parametric working models. *arXiv preprint*, 2404.16166, 2024.

[63] F. Su, W. Mou, P. Ding, and M. J. Wainwright. When is the estimated propensity score better? high-dimensional analysis and bias correction. *arXiv preprint*, 2303.17102, 2023.

[64] S. Wager. Causal inference: A statistical learning approach. *arXiv preprint*, 2024.

[65] Y. Wang, D. Kifer, and J. Lee. Differentially private confidence intervals for empirical risk minimization. *Journal of Privacy and Confidentiality*, 9(1), 2019.

[66] Y. Wang, M. Schröder, D. Frauen, J. Schweisthal, K. Hess, and S. Feuerriegel. Constructing confidence intervals for average treatment effects from multiple datasets. In *International Conference on Learning Representations (ICLR)*, 2025.

[67] L. Yao, P. Y. Li, and J. Lu. Privacy-preserving quantile treatment effect estimation for randomized controlled trials. *arXiv preprint*, 2401.14549, 2024.

[68] L. Zhang, K. K. Thekumparampil, S. Oh, and N. He. Bring your own algorithm for optimal differentially private stochastic minimax optimization. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2022.

[69] M. Zhelonkin, M. G. Genton, and E. Ronchetti. On the robustness of two-stage estimators. *Statistics & Probability Letters*, 82(4):726–732, 2012.

A Background

A.1 Estimating the average treatment effect

Under the three standard assumptions of causal inference, i.e., positivity, consistency, and unconfoundedness [e.g., 17, 57], the ATE is point-identified as

$$\tau = \mathbb{E}[Y|A=1] - \mathbb{E}[Y|A=0] = \mathbb{E}[\mu(X,1) - \mu(X,0)]. \tag{17}$$

However, estimating $\hat{\tau}$ via the empirical average of the stated difference is only unbiased if the treatment assignment is randomized and if it does not depend on X (i.e., if for a constant $\pi(x)$). To account for a non-random treatment assignment in observational data, the *inverse propensity weighted* (IPW) estimator constructs $\hat{\tau}$ by estimating $\hat{\pi}$ in the first step and then employing the estimates to take the propensity weighted empirical average of the outcomes [56].

Yet, misspecification of $\hat{\pi}$ or $\hat{\mu}$ can lead to biased ATE estimates. As a remedy, the *augmented inverse propensity weighted* (AIPW) estimator [55] employs an orthogonal loss to fulfill the *double robustness* property. Doubly robust estimators are consistent if *either* $\hat{\mu}$ *or* $\hat{\pi}$ are correctly specified. The AIPW estimator is then given by

$$\hat{\tau} = \frac{1}{n} \sum_{i=1}^{n} \left[\hat{\mu}(X,1) - \hat{\mu}(X,0) + \frac{Y_i - \hat{\mu}(X,1)}{\hat{\pi}(X_i)} A_i - \frac{Y_i - \hat{\mu}(X,0)}{1 - \hat{\pi}(X_i)} (1 - A_i) \right]. \tag{18}$$

Hence, in contrast to the IPW estimator, the AIPW estimator is an unbiased estimator of ATE and is asymptotically normal distributed [64]. This is important as we (i) aim to consistently estimate the ATE, and (ii) can derive valid CIs due to the asymptotic normality to the estimator. In our work, we thus aim to derive differentially private CIs for the ATE estimated through the AIPW estimator.

A.2 Differential privacy

In our work, we employ *output perturbation* based on the Gaussian mechanism. Of note, output perturbation is especially suitable for our task of constructing CIs for the estimated ATE due to our three reasons: output perturbation (i) retains the adherence to the causal assumptions, (ii) is model agnostic, and (iii) preserves the ability of the AIPW estimator to address the fundamental problem of causal inference [60]. In contrast, other approaches (i.e., input, objective, and gradient perturbations) fail to fulfill at least one of the points (i)–(iii) and would thus be <u>not</u> suitable for estimating the ATE with DP guarantees in a model-agnostic way.

Our work employs two important properties of differentially private algorithms, which we state below.

Lemma A.1 (Sequential composition property [23]). *Let the mechanism M_j, $j = 1, \ldots, k$ satisfy $(\varepsilon_j, \delta_j)$-DP. Then, applying the sequence of the mechanisms M_j on the same data yields $(\sum_{j=1}^{k} \varepsilon_j, \sum_{j=1}^{k} \delta_j)$-DP guarantees.*

Lemma A.2 (Post-processing property [20]). *If a mechanism M satisfies (ε, δ)-DP, then, for any function f, it holds that $f(M)$ also satisfies (ε, δ)-DP.*

A.3 Extended related work

Uncertainty quantification for causal quantities: Many different approaches for quantifying uncertainty in causal estimates exist in the literature. Many works rely on Bayesian methods [e.g., 3, 32, 38, 35]. However, Bayesian methods require prior distributions informed by domain knowledge. Therefore, they are less robust to model misspecification and, therefore, not suitable for our model-agnostic framework. A different stream focuses on estimating the conditional distribution of the treatment effect, which then can be employed for bootstrapping to construct uncertainty intervals [46, 47, 48]. Another stream of literature provides finite-sample uncertainty guarantees based on conformal prediction methods [e.g., 45, 59]. These methods can only construct intervals for the potential outcomes, not the treatment effects. Although one could construct from conformal prediction intervals for treatment effects based on the potential outcomes, those intervals commonly tend to be very wide and, therefore, less informative in practice. Importantly, <u>none</u> of the works above can provide (ε, δ)-*differentially private* confidence intervals, which is the novelty of our work.

A.4 Doubly robust variance estimation

Standard variance estimators for the AIPW estimator are, in general, only consistent if *both* nuisance functions are correctly specified [28]. Thus, these estimators are not doubly robust and, therefore, can easily lead to *undercoverage* of the CIs in practice, a highly undesirable scenario in safety-critical applications. We thus aim to estimate $\hat{\sigma}^2_{\mathrm{AIPW}}$ through a *doubly robust* variance estimator.

A common such estimator is built upon nonparametric bootstrapping [27, 62]. However, bootstrapping requires many queries to the training data, which is highly disadvantageous for estimating the variance with DP guarantees. Further, bootstrapping is computationally expensive. This is unlike doubly robust variance estimators, which have been proposed only for specific applications, namely, survey data with missing data imputation [e.g., 30, 42].

In our work, we estimate $\hat{\sigma}^2_{\mathrm{AIPW}}$ via the *empirical sandwich estimator* [36]. This has the following benefits: (i) the empirical sandwich estimator is doubly robust, and (ii) it consistently estimates the variance of the AIPW estimator [62].

Let the estimator $\hat{\theta}$ for θ_0 on data $\{z_i \mid i = 1, \ldots, n\}$ be the solution to

$$\frac{1}{n} \sum_{i=1}^{n} \psi(z_i, \hat{\theta}) = 0. \tag{19}$$

Define $\mathbf{B}(\theta) := \mathbb{E}[-\frac{\partial}{\partial \theta} \psi(Z_i, \theta)]$ with $\mathbf{B}_n(\hat{\theta}) := \frac{1}{n} \sum_{i=1}^{n} [-\frac{\partial}{\partial \hat{\theta}} \psi(Z_i, \hat{\theta})]$ and $\mathbf{M}(\theta) := \mathbb{E}[\psi(Z_i, \theta) \psi(Z_i, \theta)^T]$ with $\mathbf{M}_n(\hat{\theta}) := \frac{1}{n} \sum_{i=1}^{n} [\psi(Z_i, \hat{\theta}) \psi(Z_i, \hat{\theta})^T]$. Then, the variance of $\hat{\theta}$ can be estimated by the empirical sandwich variance estimator $\mathbf{V}_n(\hat{\theta})$ as

$$\mathbf{V}_n(\hat{\theta}) = \mathbf{B}_n(\hat{\theta})^{-1} \mathbf{M}_n(\hat{\theta}) (\mathbf{B}_n(\hat{\theta})^{-1})^T. \tag{20}$$

For estimating the variance of the AIPW estimator, it suffices to assess the estimating equation $\psi_\tau(z_i, \hat{\tau}) = \Gamma_{\hat{\eta}}(z_i) - \hat{\tau}$. It is easy to see that $\mathbf{B}_n(\hat{\tau})^{-1} = -1$ and $\mathbf{M}_n(\hat{\tau})^{-1} = (\Gamma_{\hat{\eta}}(z_i) - \hat{\tau})^2$, yielding the estimator

$$\mathbf{V}_n(\hat{\tau}) = \frac{1}{n} \sum_{i=1}^{n} (\Gamma_{\hat{\eta}}(z_i) - \hat{\tau})^2 = \frac{1}{n} \sum_{i=1}^{n} \mathrm{IF}^{\mathrm{AIPW}}(z_i, \tau_0; P)^2 \tag{21}$$

stated in Section 4.3.

B Proofs

B.1 Proofs of the supporting lemmas

B.1.1 Proof of Lemma 4.2

Lemma 4.2. The IF of the AIPW learner is given by the influence function of the final mean estimation in the second step. Specifically, the influence stemming from training the nuisance estimators π and μ can be neglected during the privatization step if the estimators are sufficiently stable.

Proof. Our proof makes use of results for two-stage influence functions [69] and influence functions of two-stage Neyman-orthogonal learners [60].

Let a two-stage estimator with first- and second-stage score functions ψ_1 and ψ_2 and a function h continuously piecewise differentiable in the second variable be defined as solving the equations

$$\mathbb{E}_F[\psi_1(z^{(1)}, S(P))] = 0 \qquad \text{and} \qquad \mathbb{E}_F[\psi_2(z^{(2)}, h(z^{(1)}, S(P)), T(P))] = 0, \qquad (22)$$

where $z_i = (z_i^{(1)}, z_i^{(2)})$ represents the data employed for first and second stage estimation, respectively, where P denotes the distribution function of z_i, and where S and T denote the functional of the first and second stage estimation, respectively. Then, the influence function of the complete two-stage estimator is given by

$$\mathrm{IF}(z, T, P) = M^{-1}\Big(\psi_2(z^{(2)}, h(z^{(1)}, S(P)), T(P)) \qquad (23)$$

$$+ \int \frac{\partial}{\partial \theta}\psi_2(\tilde{z}^{(2)}, \theta, T(P))\frac{\partial}{\partial \nu}h(\tilde{z}^{(1)}, \nu)\,\mathrm{d}F(\tilde{z}) \cdot \mathrm{IF}(z, S, P)\Big) \qquad (24)$$

where $M = -\int \frac{\partial}{\partial \xi}\psi_2(\tilde{z}^{(2)}, h(\tilde{z}^{(1)}, S(P)), \xi)\,\mathrm{d}F(\tilde{z})$ and $\mathrm{IF}(z, S, P)$ where denotes the IF of the first-stage estimation [69].

For the AIPW estimator, we have $h(\cdot) = \Gamma_{\hat{\eta}}(z)$, $T(P) = \tau$, and, with a slight abuse of notation, $S(P) = (\mu, \pi)$. We aim to show that (i) $M = 1$ and (ii) $\int \frac{\partial}{\partial \theta}\psi_2(\tilde{z}^{(2)}, \theta, T(P))\frac{\partial}{\partial \nu}h(\tilde{z}^{(1)}, \nu)\,\mathrm{d}F(\tilde{z}) = 0$.

To show (i), we observe that $\psi_2 = \frac{1}{n}\sum_{i=1}^{n}\Gamma_{\hat{\eta}}(z_i) - \tau$. Therefore, we directly get the desired equality $M = 1$. To show (ii), we observe that, for the AIPW estimator, the integral equals $\mathbb{E}[\frac{\partial}{\partial \nu}h(\tilde{z}^{(1)}, \nu)] = \mathbb{E}[\frac{\partial}{\partial \eta}\Gamma_{\hat{\eta}}(z)]$, where $\eta = (\hat{\mu}, \hat{\pi})$. Similar to Schröder et al. [60], we now observe that, due to the Neyman-orthogonality of the AIPW estimator [e.g., 64] and the stability of the nuisance estimators, we have $\mathbb{E}[\frac{\partial}{\partial \eta}\Gamma_{\hat{\eta}}(z)] = 0$, which yields the desired result. □

B.1.2 Proof of Lemma 4.4

Lemma 4.4. The influence function of $\hat{\sigma}_{\mathrm{AIPW}}^2$ is given by

$$\mathrm{IF}^{\sigma}(z, \sigma_0, P) = (\Gamma_{\hat{\eta}}(z) - \hat{\tau})^2 - \hat{\sigma}_{\mathrm{AIPW}}^2. \qquad (25)$$

Proof. Recall that we estimate $\hat{\sigma}_{\mathrm{AIPW}}^2$ by

$$\hat{\sigma}_{\mathrm{AIPW}}^2 = \frac{1}{n}\sum_{i=1}^{n}\mathrm{IF}^{\mathrm{AIPW}}(z_i, \tau_0; P)^2 \qquad (26)$$

with $\mathrm{IF}^{\mathrm{AIPW}}(z, \tau_0; P) = \Gamma_{\hat{\eta}}(z) - \hat{\tau}$. Therefore, $\hat{\sigma}_{\mathrm{AIPW}}^2$ is estimated as the mean of the variable $(\Gamma_{\hat{\eta}}(z_i) - \hat{\tau})^2$. The result then directly follows with the IF of the mean. □

B.2 Proofs of the main theorems

B.2.1 Proof of Theorem 4.3

Theorem 4.3. Let $z := (a, x, y)$ define a data sample following the joint distribution $\mathcal{Z}$ and $\hat{\eta} = (\hat{\mu}, \hat{\pi})$ the estimated nuisance functions. Furthermore, let D be the training dataset with $|D| = n$. Define

$$\hat{\tau}_{\text{DP}} := \hat{\tau} + \sup_{z \in \mathcal{Z}} \left\| \Gamma_{\hat{\eta}}(z) - \hat{\tau} \right\| \cdot \frac{5 \sqrt{2 \ln(n) \ln(2/\delta)}}{\varepsilon n} \cdot U, \tag{27}$$

where $U \sim \mathcal{N}(0, 1)$. Then, $\hat{\tau}_{\text{DP}}$ is (ε, δ)-differentially private.

Proof. First, recall that the influence function of the AIPW estimator is given by $\text{IF}^{\text{AIPW}}(z, \tau_0; P) = \Gamma_{\hat{\eta}}(z) - \hat{\tau}$. Therefore, $\gamma_\tau := \sup_{z \in \mathcal{Z}} \left\| \Gamma_{\hat{\eta}}(z) - \hat{\tau} \right\|$ denotes the gross-error sensitivity of the AIPW estimator.

Our proof builds upon the idea of upper-bounding the global sensitivity $\Delta_2(\mathbf{f})$ required for privatization with that Gaussian mechanism (Definition 3.2) by the so-called *smooth-sensitivity* [50]. Specifically, for the ξ-smooth sensitivity $SS_\xi(f, D)$ of the estimator f with $\xi = \frac{\varepsilon}{4(d + 2 \log(2/\delta))}$, it holds that

$$\mathbf{f}_D^{\text{DP}}(\mathbf{x}) = \mathbf{f}_D(\mathbf{x}) + \frac{5 \sqrt{2 \log(2/\delta)}}{\varepsilon} SS_\xi(f, D) \cdot U, \tag{28}$$

where $U \sim \mathcal{N}(0, 1)$ is (ε, δ)-differential private with

$$SS_\xi(f, D) := \sup_{D'} \{ \exp(-\xi d_{\text{H}}(D, D')) LS(f, D') \mid D' \in \mathcal{Z}^n \}, \tag{29}$$

where $\mathcal{Z}^n$ denotes the data domain and $LS(\cdot)$ is the local sensitivity given by

$$LS(f, D) := \sup_{D'} \left\{ \| f_D - f_{D'} \| \; \middle| \; d_{\text{H}}(D, D') = 1 \right\} \tag{30}$$

[see 50]. However, similar to the global sensitivity, the smooth sensitivity is extremely expensive or even infeasible to calculate, depending on the machine learning model employed for nuisance estimation. Therefore, we follow Avella-Medina [5] and aim to upper bound the smooth sensitivity by the appropriately scaled gross-error sensitivity of the AIPW estimator.

Observe that the AIPW estimate $\hat{\tau}$ solves the equation

$$\sum_{i=1}^{n} \text{IF}^{\text{AIPW}}(z_i, \tau_0; P) = 0. \tag{31}$$

The influence function is differentiable with respect to $\hat{\tau}$ with derivative $\frac{\partial}{\partial \hat{\tau}} \text{IF}^{\text{AIPW}}(z, \tau_0; P) = 1$. Furthermore, since we assume the data to stem from a bounded domain, there exists a constant $K \in \mathbb{R}$ such that, for the gross-error sensitivity of $\hat{\tau}$, one has $\gamma_\tau \leq K$. As a result, we can upper bound the smooth sensitivity of the AIPW estimator by $SS_\xi(\hat{\tau}, D) \leq \frac{\sqrt{\log(n)}}{n} \gamma_\tau$ [5], leading to the desired result in Theorem 4.3. $\square$

B.2.2 Proof of Theorem 4.5

Theorem 4.5. Let $z := (a, x, y)$ define a data sample following the joint distribution $\mathcal{Z}$, and let $\hat{\eta} = (\hat{\mu}, \hat{\pi})$ be the estimated nuisance functions. Furthermore, let D be the training dataset with $|D| = n$. For $U \sim \mathcal{N}(0, 1)$, we define

$$\hat{\sigma}_{\text{DP}}^2 := \max \left\{ 0, \hat{\sigma}_{\text{AIPW}}^2 + \sup_{z \in \mathcal{Z}} \left\| \text{IF}^\sigma(z, \sigma_0, P) \right\| \cdot \frac{5 \sqrt{2 \ln(n) \ln(2/\delta)}}{\varepsilon n} \cdot U \right\}. \tag{32}$$

Then, $\hat{\sigma}_{\text{DP}}^2$ is (ε, δ)-differentially private.

Proof. We proceed in the same manner as in the proof of Theorem 4.3. We observe that $\hat{\sigma}^2_{\text{AIPW}}$ solves $\sum_{i=1}^{n} \text{IF}^{\sigma}(z_i, \sigma_0; P) = 0$ with $\frac{\partial}{\partial \hat{\sigma}^2} \text{IF}^{\sigma}(z, \sigma_0; P) = 1$. Again, since we assume the data to stem from a bounded domain, there exists a constant $K \in \mathbb{R}$ such that, for the gross-error sensitivity of $\hat{\sigma}^2$, we have $\gamma_\sigma := \sup_{z \in \mathcal{Z}} \left\| \text{IF}^{\sigma}(z, \sigma_0, P) \right\| \leq K$. Then, with Theorem 1 in Avella-Medina [5], we can upper bound the smooth sensitivity $SS_\xi(\hat{\sigma}^2, D)$ by $\frac{\sqrt{\log(n)}}{n} \gamma_\sigma$. As a result, we get that the quantity

$$\hat{\sigma}^2_{\text{AIPW}} + \sup_{z \in \mathcal{Z}} \left\| \text{IF}^{\sigma}(z, \sigma_0, P) \right\| \cdot \frac{5\sqrt{2\ln(n)\ln(2/\delta)}}{\varepsilon n} \cdot U \tag{33}$$

is (ε, δ)-private. Now, observe that the maximum function does not require further queries to the underlying data. Therefore, by the post-processing theorem of DP, $\hat{\sigma}^2_{\text{DP}}$ is as well (ε, δ)-private. $\square$

B.2.3 Proof of Theorem 4.6

Theorem 4.6. Let $z := (a, x, y)$ define a data sample following the joint distribution $\mathcal{Z}$, and let $\hat{\eta} = (\hat{\mu}, \hat{\pi})$ be the estimated nuisance functions. We define the privacy budget (ε, δ) and the sub-budgets $\varepsilon_1, \varepsilon_2, \delta_1, \delta_2$, for ATE and variance estimation under DP such that $\varepsilon_1 + \varepsilon_2 = \varepsilon$ and $\delta_1 + \delta_2 = \delta$. Furthermore, let D be the training dataset with $|D| = n$, and let $\hat{\tau}_{\text{DP}}$ denote the $(\varepsilon_1, \delta_1)$-differentially private ATE estimate and $\hat{\sigma}_{\text{DP}}$ denote the $(\varepsilon_2, \delta_2)$-differentially private variance estimate. Then, the valid and (ε, δ)-differentially private CI is given by

$$\text{CI}_{\text{DP}} := \left[\hat{\tau}^{\text{DP}} \pm \Phi^{-1} \left(1 - \frac{\alpha}{2} \right) \sqrt{\frac{\hat{V}_{\text{DP}}}{n}} \right], \tag{34}$$

where $\Phi^{-1}(1 - \alpha/2)$ denotes the respective quantile of the standard normal distribution, γ_τ is the gross-error sensitivity of the ATE estimation, and

$$\hat{V}_{\text{DP}} = \hat{\sigma}^2_{\text{DP}} + \gamma_\tau^2 \cdot \frac{50\ln(n)}{n\varepsilon_1^2} \ln\left(\frac{2}{\delta_1} \right). \tag{35}$$

Proof. We need to show that (i) the provided CI is differentially private and (ii) that it is valid in the sense that the CI has asymptotic coverage at the targeted confidence level $1 - \alpha$.

First, we show (i): By Theorem 4.3 and Theorem 4.5, $\hat{\tau}_{\text{DP}}$ and $\hat{\sigma}_{\text{DP}}$ are $(\varepsilon_1, \delta_1)$- and $(\varepsilon_2, \delta_2)$-differentially private, respectively. As $\hat{V}_{\text{DP}}$ does not require further access to single data points, it as well is $(\varepsilon_2, \delta_2)$-differentially private by the post-processing property of DP (see Lemma A.2). As CI_{DP} states a non-data dependent combination of two privatized estimates, , CI_{DP} is $(\varepsilon_1 + \varepsilon_2, \delta_1 + \delta_2)$-differentially private by the parallel decomposition and the post-processing properties of DP (see Lemmas A.1 and A.2).

Now, we turn to proving the validity of the CI (ii): Recall from Section 4.3 that

$$\sqrt{n} \left(\hat{\tau}_{\text{DP}} - \tau_0 \right) \rightarrow \sqrt{\sigma^2_{\text{AIPW}}} \cdot \mathcal{N}(0, 1). \tag{36}$$

As discussed in Section 4.3, we also aim to account for the finite sample variance stemming from privatization, to make the intervals not only asymptotically valid but also useful in finite sample settings. Thus, a valid CI which accounts for the finite sample privatization noise could be constructed around $\hat{\tau}_{\text{DP}}$ with variance $\sigma^2 = \sigma^2_{\text{AIPW}} + n \, r(\varepsilon, \delta, n)^2$. However, as σ^2_{AIPW} is non-private, the resulting CI would also not be private.

Now observe that, by construction of σ^2_{DP}, it also holds that $\sqrt{n} \left(\sigma^2_{\text{DP}} - \sigma^2_{\text{AIPW}} \right) \rightarrow 0$, which proves the validity of CI_{DP}. $\square$

C Experiments

C.1 Datasets

C.1.1 Synthetic data

We generate our datasets similar to Oprescu et al. [53]:

$$X_i \sim \mathcal{U}[0,1]^p \tag{37}$$

$$A_i = \mathbf{1}\{(X^T\beta)_i \geq \eta_i\} \tag{38}$$

$$Y_i = \tau A_i + X_i^T\gamma + \epsilon_i \tag{39}$$

where $\epsilon_i, \eta_i \sim \mathcal{U}[-1,1]$ are independent noise terms. The coefficients β and γ are sparse vectors with random support size s, where non-zero entries are drawn uniformly with $\beta_j \sim \mathcal{U}[0, 0.3]$ and $\gamma_j \sim \mathcal{U}[0,1]$ for j in the support. To satisfy the overlap assumption, propensity scores are clipped in the range $(0.1, 0.9)$, whereby the true average treatment effect τ is set to 1.0. Here, we consider two datasets: a low-dimensional dataset with $p = 2$ covariates (Dataset 1) and a high-dimensional dataset with $p = 24$ and $s = 6$ (Dataset 2).

C.1.2 Medical data

The RHC dataset contains detailed real-world observations of 5735 patients admitted to intensive care units at different hospitals as part of the Study to Understand Prognoses and Preferences for Outcomes and Risks of Treatments (SUPPORT). We aim to predict the average treatment effect of receiving a Swan-Ganz catheterization, a diagnostic procedure performed on a pulmonary artery, on the probability of death in the 180 days after hospitalization. Our estimation is based on 8 confounders from medical practice (e.g., primary disease, heart rate, Glasgow coma score) and a binary treatment (Swan-Ganz catheterization in the first 24 hours after admission). We use a standard preprocessing pipeline that imputes missing values and removes outliers; our preprocessing pipeline is available in our GitHub repository.

C.2 Implementation details

All of our experiments are implemented in Python. We provide our code in our GitHub repository: `https://github.com/m-schroder/PrivATE`. The experiments were run on an AMD Ryzen 7 PRO 6850U 2.70 GHz CPU with eight cores and 32GB RAM. The runtime varied for the different experiments (longer runtime for the more complex dataset 2). However, all experiments are easily feasible to compute on all standard computing resources within a short time.

Our framework is highly flexible and can be used with any base machine learning method. We implement two versions: (1) a regression-based learner using logistic regression for the propensity score and kernel ridge regression (RBF kernel, $\alpha = 0.1$) for the outcome model, and (2) a neural network-based learner with a simple multi-layer perceptron using one hidden layer of 32 neurons and $\tanh$ activation. The neural networks are optimized via stochastic gradient descent and regularized via L2-regularization ($\alpha = 0.1$).

The PRIVATE framework includes the calculation of a supremum over a bounded data space. We implemented this maximization problem using L-BFGS-B [14], a limited-memory Broyden–Fletcher–Goldfarb–Shanno algorithm for solving nonlinear optimization problems with bounded variables. We use default hyperparameters and use $n = 10$ random starting points for every optimization.

By default, we choose $\varepsilon_1 = \varepsilon_2$ as well as $\delta_1 = \delta_2$ if not specified otherwise.

D   Trade-off between privacy and utility of the CIs

Introducing privacy to the ATE estimation and the CIs lowers the informativeness of the intervals due to the additional noise and thus wider intervals. However, this is expected and inherent to privatization, and thus <u>not</u> specific to our framework. In practice, it might be interesting to visualize the privacy-utility tradeoff based on a user-defined utility function to decide on a level of privatization for a new analysis.

We performed such an analysis for our experiments, by comparing the utility of PRIVATE against the utility of a bootstrapping method on the private ATE and the non-private estimation. Utility is defined as a weighted sum of privacy and interval width. We present the results for various privacy budgets in Fig. 5. For all utility weightings, we observe that our method outperforms the baselines.
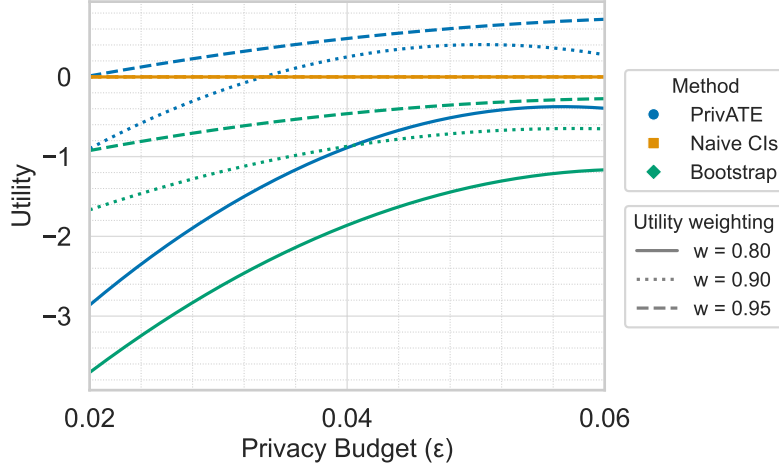

Figure 5: **Privacy-utility trade-off.** We report the utility curves with respect to the privacy budget for various utility functions (confidence level 0.95). As expected, we observe our PRIVATE to outperform the naive and the bootstrapping method with increasing utility for increasing weight on the privacy constraint. The utility of the naive CIs does not vary significantly for different utility weightings.

E Extension to interventional RCT data

We now demonstrate the applicability of PRIVATE to experimental data. In doing so, we can now additionally compare our method against the private differences of means method by Guha and Reiter [29], yet which is restricted to experimental data (and thus <u>not</u> applicable to the observational data used throughout our main paper). Again, we evaluate PRIVATE using simulated and real-world RCT data.

E.1 Dataset and implementation details

Each sample consists of 1 uniformly distributed confounder, as in Dataset 1. To simulate an RCT setting, the treatment was then assigned by complete randomization with a factor of 0.5. The outcome generation again resembles the one from Dataset 1. To create the dataset, we sampled 1000 data samples from the described data-generating process. The implementation remained the same as for the other experiments.

E.2 Results

In Table 3, we present the average coverage and the corresponding standard deviation for both methods for confidence levels $\{0.80, 0.90, 0.95\}$ across ten runs. As expected, both methods roughly achieve the desired coverage.

Confidence	Empirical coverage		
$1 - \alpha$	Difference of means	PRIVATE (Kernel)	PRIVATE (NN)
0.80	0.820 ± 0.039	0.780 ± 0.042	0.880 ± 0.033
0.90	0.920 ± 0.027	0.900 ± 0.030	0.890 ± 0.031
0.95	0.950 ± 0.022	0.940 ± 0.024	0.970 ± 0.017

Table 3: **Synthetic RCT evaluation.**

When comparing the methods in terms of the widths of the intervals, we observe a significant difference (see Fig. 6): The intervals given by the private differences of means method are much wider than the ones returned by our PRIVATE and thus less informative. In other words: *our method is preferred*.
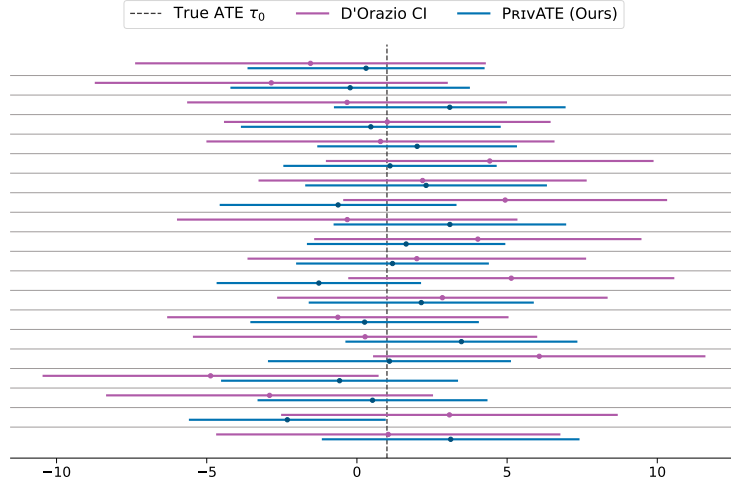

Figure 6: **Visualization of the CIs on RCT data.** The CIs are reported for both PRIVATE and the difference in means method by Guha and Reiter [29]. For both, we use the NN base learner across different runs with privacy budgets $\varepsilon = 0.5$ and $\delta = 10^{-5}$. $\Rightarrow$ The intervals of the difference in means method are significantly wider than the ones of PRIVATE. Our method returns more informative intervals and is thus preferred.