Causal Inference with Differential Privacy

Sharmistha Guha

Assistant Professor, Department of Statistics, Texas A&M University,

3143 TAMU, College Station, TX 77843-3143, E-mail: sharmistha@tamu.edu

Jerome P. Reiter

Professor, Department of Statistical Science, Duke University,

214 Old Chemistry Building, Durham, NC 27708-0251, E-mail: jreiter@duke.edu

March 21, 2024

Abstract

In the social and health sciences, researchers often make causal inferences using sensitive variables. These researchers, as well as the data holders themselves, may be ethically and perhaps legally obligated to protect the confidentiality of study participants' data. It is now known that releasing any statistics, including estimates of causal effects, computed with confidential data leaks information about the underlying data values. Thus, analysts may desire to use causal estimators that can provably bound this information leakage. Motivated by this goal, we develop algorithms for estimating weighted average treatment effects with binary outcomes that satisfy the criterion of differential privacy. We present theoretical results on the accuracy of several differentially private estimators of weighted average treatment effects. We illustrate the empirical performance of these estimators using simulated data and a causal analysis using data on education and income.

Keywords: Causal; Confidentiality; Observational; Privacy; Propensity.

1 Introduction

Many causal studies measure sensitive binary outcome variables that data stewards are ethically, and sometimes even legally, obligated to keep confidential. As hypothetical but realistic examples, the outcomes could be whether or not a patient is cured of a stigmatized disease after treatment, a student passes or fails a test after receiving an intervention, or a person is employed or unemployed after job training. In each of these cases, study participants would not want the data analyst to release information in a manner that reveals their individual outcomes. Furthermore, causal studies typically include additional sensitive or identifying variables that analysts want to use as covariates; these too may be confidential.

Data stewards routinely put controls in place to reduce risks of unintended disclosures. For example, often they restrict access to the confidential data to vetted data analysts. However, researchers in data privacy have shown that every statistic computed with confidential data leaks information about that data (Dwork and Roth, 2014; Dwork *et al.*, 2017). Given enough released information of sufficient accuracy, ill-intentioned users may be able to learn confidential information. Thus, data stewards and analysts may seek to bound the information leakage when sharing results of confidential data analysis.

One way to do so is to require methods to provide formal guarantees of confidentiality protection for any data release. Among such methods, algorithms that satisfy differential privacy (Dwork, 2006) have become a gold standard. Differential privacy is a mathematical criterion that encodes the idea that the released statistic should not be overly sensitive to the presence or absence of any particular individual; see Section 2.2 for details. Researchers have developed differentially private algorithms for a variety of estimation tasks, including significance tests (e.g., Barrientos *et al.*, 2019; Balle *et al.*, 2020; Pensia *et al.*, 2023), regression (e.g., Zhang *et al.*, 2012; Wang *et al.*, 2015; Fang *et al.*, 2019; Gaboardi *et al.*, 2019), and machine learning (e.g., Mivule *et al.*, 2012; Ji *et al.*, 2014; Abadi *et al.*, 2016; Triastcyn and Faltings, 2020; Zheng *et al.*, 2020; Blanco-Justicia *et al.*, 2022), among many others.

The literature includes few approaches to differentially private causal inference, particularly in observational studies. D'Orazio *et al.* (2015) present differentially private algorithms for estimating the differences of two means in matched pairs designs, which are common

in causal inference. Lee *et al.* (2019) construct a differentially private inverse-probability weighting treatment effect estimator. They first fit a differentially private propensity score model to determine the weights, and then add Gaussian noise under (ϵ, δ) differential privacy to perturb the resulting weighted treatment effect estimate. Their method does not provide standard errors or interval estimates for the treatment effect. Niu *et al.* (2022) use partitions of the data to estimate parts of machine learning algorithms that are used for causal inference, and combine all the parts together at the end to arrive at a causal estimate. Their method also does not provide standard errors or interval estimates. Finally, Ohnishi and Awan (2023) show that plugging-in differentially private versions of parts of causal estimators can result in biased estimates. They also provide a Bayesian version of causal inference under the local differential privacy model, i.e., when individuals perturb their own data before providing it to a central party that does computations.

In this article, we contribute to this literature by proposing differentially private algorithms for causal inference with binary outcomes. The algorithms can be used with a variety of weighted average treatment effect estimators. Unlike other approaches, they generate standard errors and confidence intervals for these estimators. The basic idea is to split the data into M disjoint groups, estimate causal effects and standard errors in each partition, aggregate the results, and add differentially private noise to the aggregated results. We illustrate the approach using simulation studies and an analysis of data from the 1994 U.S. census in a study of the effect of education on earnings.

The remainder of this article is organized as follows. In Section 2, we review key concepts from causal inference and differential privacy. In Section 3, we present the differentially private treatment effect point and interval estimators. In Section 4, we present results of simulation studies showing the performance of the proposed methodology in various scenarios. In Section 5, we illustrate the methodology using the 1994 U.S. census data. Finally, in Section 6, we conclude with a discussion.

2 Review of Causal Inference and Differential Privacy

In Section 2.1, we introduce the weighted average treatment effect (WATE) and methods for estimating the WATE. In Section 2.2, we review differential privacy and several algorithms

that satisfy it. Throughout the article, we suppose sample sizes to be large enough that large sample approximations to sampling distributions are empirically valid.

2.1 Overview of WATE Estimation

We use the potential outcome framework for causal inference (Rubin, 1974). Let $z = 1$ and $z = 0$ indicate assignment to the treatment and control conditions, respectively. Let y be an outcome variable. We seek to learn the causal effect of z on y. For any unit in the study population, we conceive of two potential outcomes, $y(1)$ and $y(0)$, corresponding to the outcome measured when $z = 1$ and $z = 0$, respectively. For any unit, we observe only one of $y(1)$ and $y(0)$, which we write as $y = zy(1) + (1 - z)y(0)$. We consider $y(0)$ and $y(1)$ as binary outcomes, i.e., $y(0), y(1) \in \{0, 1\}$. We assume the stable unit treatment value assumption (SUTVA) which contains two sub-assumptions, no interference between units (i.e., the treatment applied to one unit does not affect the outcome for another unit) and no different versions of a treatment (Rubin, 1974). We also define the $p \times 1$ vector of covariates $\boldsymbol{x}$, which are variables unaffected by treatment assignment z. We assume that $P(z = 1|\boldsymbol{x}) > 0$, i.e., the probability of assigning treatment is positive for every unit. Finally, we assume strong ignorability (Rosenbaum and Rubin, 1983) so that the vector of potential outcomes $(y(0), y(1))$ is independent of z given $\boldsymbol{x}$.

Many causal inference procedures utilize propensity scores $P(z = 1|\boldsymbol{x})$, i.e., the probability of assignment to the treatment group given the covariates $\boldsymbol{x}$. As shown by Rosenbaum and Rubin (1983), the treatment assignment is independent of $\boldsymbol{x}$ given $P(z = 1|\boldsymbol{x})$ under SUTVA and strong ignorability. Propensity scores are typically estimated using binary regressions of z on $\boldsymbol{x}$. These estimated scores are used in a variety of causal estimators, especially in defining weighted sums of the outcomes as treatment effect estimators, as we use here. In what follows, we refer to estimated propensity scores using $e(\boldsymbol{x})$.

To compare outcomes under treatment and control, we define the conditional average controlled difference for a given $\boldsymbol{x}$,

$$\tau(\boldsymbol{x}) = E[y|z = 1, \boldsymbol{x}] - E[y|z = 0, \boldsymbol{x}]. \tag{1}$$

Under strong ignorability, $E[y(z)|\boldsymbol{x}] = E[y|\boldsymbol{x}, z]$, so that $\tau(\boldsymbol{x})$ in (1) becomes the average

treatment effect conditional on $\boldsymbol{x}$, i.e., $\tau(\boldsymbol{x}) = E[y(1) - y(0)|\boldsymbol{x}]$. Typically, the (potential) outcomes are compared not for a single $\boldsymbol{x}$; rather, they are averaged over a hypothesized target distribution of the covariates. The choice of the distribution corresponds to the region of the covariate space for the target population of interest. For example, if one seeks to estimate the effect of the treatment on the treated, the relevant covariate distribution is that of the treated cases.

Let the marginal density of $\boldsymbol{x}$ be $f(\boldsymbol{x})$, defined with respect to a base measure $\Delta(\boldsymbol{x})$ (a product of counting measure for categorical variables and Lebesgue measure for continuous variables). For many common target populations in causal inference, the distribution of the covariates can be represented as $g(\boldsymbol{x}) = f(\boldsymbol{x})t(\boldsymbol{x})$, where $t(\cdot)$ is a pre-specified function of $\boldsymbol{x}$. Using this expression, we define a general class of estimands by the expectation of the conditional average controlled difference over the target population,

$$\tau = \frac{\int \tau(\boldsymbol{x})t(\boldsymbol{x})f(\boldsymbol{x})\Delta(d\boldsymbol{x})}{\int t(\boldsymbol{x})f(\boldsymbol{x})\Delta(d\boldsymbol{x})}. \tag{2}$$

The class of estimators defined in (2) is referred to as the weighted average treatment effect (WATE) for causal comparisons (Hirano *et al.*, 2003). Specification of $t(\cdot)$ defines the target population and WATE estimands. Here, we consider three different WATEs. When $t(\boldsymbol{x}) = 1$, the corresponding target population is the combined (treated and control) population, and the estimand is the average treatment effect (ATE). When $t(\boldsymbol{x}) = e(\boldsymbol{x})$, the target population is the treated subpopulation, and the estimand is the average treatment effect for the treated (ATT). Finally, when $t(\boldsymbol{x}) = 1 - e(\boldsymbol{x})$, the target population is the control subpopulation, and the estimand is the average treatment effect for the control (ATC).

For any unit i in a study where $i = 1, \ldots, n$, let their covariates be $\boldsymbol{x}_i$, their treatment status z_i, and their outcome $y_i = z_i y_i(1) + (1 - z_i)y_i(0)$. The observed data are $\boldsymbol{D} = \{(y_i, \boldsymbol{x}_i, z_i) : i = 1, \ldots, n\}$. We refer to the sampled covariate values as $\boldsymbol{X} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$. For $i = 1, \ldots, n$, let $w_{1i} = t(\boldsymbol{x}_i)/e(\boldsymbol{x}_i)$, and let $w_{0i} = t(\boldsymbol{x}_i)/(1-e(\boldsymbol{x}_i))$. A consistent estimator

τ	$(w_{0i},\, w_{1i})$	Estimator	Estimated variance
ATE	$\left(\frac{1}{1-e(\boldsymbol{x}_i)},\, \frac{1}{e(\boldsymbol{x}_i)}\right)$	$\frac{\sum_{i=1}^n \frac{z_i y_i}{e(\boldsymbol{x}_i)}}{\sum_{i=1}^n \frac{z_i}{e(\boldsymbol{x}_i)}} - \frac{\sum_{i=1}^n \frac{(1-z_i) y_i}{1-e(\boldsymbol{x}_i)}}{\sum_{i=1}^n \frac{(1-z_i)}{1-e(\boldsymbol{x}_i)}}$	$\frac{\sum_{i=1}^n \left\{ \frac{v_1(\boldsymbol{x}_i)}{e(\boldsymbol{x}_i)} + \frac{v_0(\boldsymbol{x}_i)}{1-e(\boldsymbol{x}_i)} \right\}}{n^2}$
ATT	$\left(\frac{e(\boldsymbol{x}_i)}{1-e(\boldsymbol{x}_i)},\, 1\right)$	$\frac{\sum_{i=1}^n z_i y_i}{\sum_{i=1}^n z_i} - \frac{\sum_{i=1}^n \frac{(1-z_i) y_i e(\boldsymbol{x}_i)}{1-e(\boldsymbol{x}_i)}}{\sum_{i=1}^n \frac{(1-z_i) e(\boldsymbol{x}_i)}{1-e(\boldsymbol{x}_i)}}$	$\frac{\sum_{i=1}^n e(\boldsymbol{x}_i)^2 \left\{ \frac{v_1(\boldsymbol{x}_i)}{e(\boldsymbol{x}_i)} + \frac{v_0(\boldsymbol{x}_i)}{1-e(\boldsymbol{x}_i)} \right\}}{\left[\sum_{i=1}^n e(\boldsymbol{x}_i))\right]^2}$
ATC	$\left(1,\, \frac{1-e(\boldsymbol{x}_i)}{e(\boldsymbol{x}_i)}\right)$	$\frac{\sum_{i=1}^n \frac{z_i y_i (1-e(\boldsymbol{x}_i))}{e(\boldsymbol{x}_i)}}{\sum_{i=1}^n \frac{z_i (1-e(\boldsymbol{x}_i))}{e(\boldsymbol{x}_i)}} - \frac{\sum_{i=1}^n (1-z_i) y_i}{\sum_{i=1}^n (1-z_i)}$	$\frac{\sum_{i=1}^n (1-e(\boldsymbol{x}_i))^2 \left\{ \frac{v_1(\boldsymbol{x}_i)}{e(\boldsymbol{x}_i)} + \frac{v_0(\boldsymbol{x}_i)}{1-e(\boldsymbol{x}_i)} \right\}}{\left[\sum_{i=1}^n (1-e(\boldsymbol{x}_i))\right]^2}$

Table 1: The expressions for w_{0i}, w_{1i}, the estimated treatment effect and its estimated variance for different choices of target population represented by $t(\boldsymbol{x}_i)$. ATE is the average treatment effect for everyone ($t(\boldsymbol{x}) = 1$). ATT is the average treatment effect for the treated ($t(\boldsymbol{x}) = e(\boldsymbol{x})$). ATC is the average treatment effect for the controls ($t(\boldsymbol{x}) = 1 - e(\boldsymbol{x})$).

of τ for any target population represented by the function $t(\cdot)$ is given by

$$\hat{\tau} = \frac{\sum_{i=1}^n w_{1i} z_i y_i}{\sum_{i=1}^n w_{1i} z_i} - \frac{\sum_{i=1}^n w_{0i} (1 - z_i) y_i}{\sum_{i=1}^n w_{0i} (1 - z_i)}. \tag{3}$$

Expressions for w_{0i}, w_{1i} and $\hat{\tau}$ corresponding to the ATE, ATT and ATC are given in Table 1. We denote the treatment effect estimators as $\hat{\tau}_{ATE}, \hat{\tau}_{ATT}, \hat{\tau}_{ATC}$, respectively.

For any of these estimators, which we write generically as $\hat{\tau}$, we can approximate the variance $V[\hat{\tau}]$ using the decomposition, $V[\hat{\tau}] = E_{\boldsymbol{x}} V[\hat{\tau}|\boldsymbol{X}] + V_{\boldsymbol{x}} E[\hat{\tau}|\boldsymbol{X}]$. Li *et al.* (2018) derive the component of variation due to residual (model) variation conditional on $\boldsymbol{X}$. Specifically, if $\mathrm{V}[y(1)|\boldsymbol{X}] = v_1(\boldsymbol{x})$ and $\mathrm{V}[y(0)|\boldsymbol{X}] = v_0(\boldsymbol{x})$ denote the variances of the outcome given the covariates for the treated and control groups, respectively, Li *et al.* (2018) show that $E_{\boldsymbol{x}} V[\hat{\tau}|\boldsymbol{X}]$ can be approximated by

$$V = \frac{1}{n} \int t(\boldsymbol{x})^2 \left\{ \frac{v_1(\boldsymbol{x})}{e(\boldsymbol{x})} + \frac{v_0(\boldsymbol{x})}{(1 - e(\boldsymbol{x}))} \right\} f(\boldsymbol{x}) \Delta(\boldsymbol{x}) / \left\{ \int t(\boldsymbol{x}) f(\boldsymbol{x}) \Delta(\boldsymbol{x}) \right\}^2, \tag{4}$$

when the sample size n is large. Imbens (2004) shows that $E_{\boldsymbol{x}} V[\hat{\tau}|\boldsymbol{X}]$ is typically much larger than $V_{\boldsymbol{x}} E[\hat{\tau}|\boldsymbol{X}]$. Therefore, the general strategy is to approximate $V[\hat{\tau}]$ by (4). The

expression in (4) can be empirically approximated by,

$$\hat{V} = \frac{\frac{1}{n}\left[\frac{1}{n}\sum_{i=1}^{n} t(\boldsymbol{x}_i)^2 \left\{\frac{v_1(\boldsymbol{x}_i)}{e(\boldsymbol{x}_i)} + \frac{v_0(\boldsymbol{x}_i)}{1-e(\boldsymbol{x}_i)}\right\}\right]}{\left[\frac{1}{n}\sum_{i=1}^{n} t(\boldsymbol{x}_i))\right]^2} = \frac{\left[\sum_{i=1}^{n} t(\boldsymbol{x}_i)^2 \left[\frac{v_1(\boldsymbol{x}_i)}{e(\boldsymbol{x}_i)} + \frac{v_0(\boldsymbol{x}_i)}{1-e(\boldsymbol{x}_i)}\right]\right]}{\left[\sum_{i=1}^{n} t(\boldsymbol{x}_i))\right]^2}. \tag{5}$$

The estimated variance $\hat{V}$ corresponding to $\hat{\tau}_{ATE}, \hat{\tau}_{ATT}$ and $\hat{\tau}_{ATC}$ are denoted $\hat{V}_{ATE}$, $\hat{V}_{ATT}$ and $\hat{V}_{ATC}$, respectively. Their expressions are in Table 1. With large n, 95% confidence intervals for τ are constructed based on a large-sample normal approximation, $(\hat{\tau} - 1.96\sqrt{\hat{V}}, \hat{\tau} + 1.96\sqrt{\hat{V}})$.

In some settings, values of $e(\boldsymbol{x}_i)$ can be close to zero or one, which can result in inflated variances. In such cases, one remedy is to replace $e(\boldsymbol{x})$ with a truncated propensity score, given by

$$e^T(\boldsymbol{x}) = \begin{cases} 1-a & \text{if } e(\boldsymbol{x}_i) > 1-a \\ e(\boldsymbol{x}) & \text{if } a \le e(\boldsymbol{x}_i) \le 1-a \\ a & \text{if } e(\boldsymbol{x}_i) < a \end{cases} \tag{6}$$

for some user-defined parameter $a > 0$. The value of a typically is chosen to be small, so that truncation affects only the few units with $e(\boldsymbol{x}_i)$ near zero or one. Inferences are based on the expressions in Table 1 replacing $e(\boldsymbol{x}_i)$ with $e^T(\boldsymbol{x}_i)$. We denote the causal estimators based on truncated propensity scores as $\hat{\tau}^T_{ATE}, \hat{\tau}^T_{ATT}$, and $\hat{\tau}^T_{ATC}$, with corresponding variance estimates $\hat{V}^T_{ATE}$, $\hat{V}^T_{ATT}$, and $\hat{V}^T_{ATC}$.

Trimming estimated propensity scores is a common strategy in the causal inference literature, especially to avoid large variance and poor finite-sample performance due to large values of w_{1i} or w_{0i} (Kang and Schafer, 2007). The idea was first discussed in medical applications (Vincent *et al.*, 2002; Grzybowski *et al.*, 2003; Kurth *et al.*, 2006) and formalized by Crump *et al.* (2009).

2.2 Differential Privacy: Overview of Key Concepts

Let $\mathcal{P}$ be a stochastic algorithm that takes a database $\boldsymbol{D}$ as input and outputs a quantity $\boldsymbol{q}$, i.e., $\mathcal{P}(\boldsymbol{D}) = \boldsymbol{q}$. We call $\boldsymbol{D}$ and $\boldsymbol{D}'$ neighboring databases if there exists only one record $\{d\} \in \boldsymbol{D}$ and one record $\{d'\} \in \boldsymbol{D}'$ such that $d \neq d'$ and $\boldsymbol{D} - \{d\} = \boldsymbol{D}' - \{d'\}$.

Definition 1 *(ϵ-Differential Privacy) An algorithm $\mathcal{P}$ satisfies $\epsilon-$differential privacy (denoted as ϵ-DP), if for any pair of neighboring databases $(\boldsymbol{D}, \boldsymbol{D}')$, and any non-negligible measurable set $S \subseteq$ range$(\mathcal{P})$, $\frac{P(\mathcal{P}(\boldsymbol{D})\in S)}{P(\mathcal{P}(\boldsymbol{D}')\in S)} \leq \exp(\epsilon)$.*

Thus, $\mathcal{P}$ satisfies ϵ-DP when the distributions of its outputs are similar for any two neighboring databases, where the factor $\exp(\epsilon)$ defines the similarity. The ϵ, known as the privacy loss budget, controls the degree of confidentiality protection provided by $\mathcal{P}$, with greater protection guarantees implied by lower values. ϵ-DP is a strong criterion, because an attacker who has access to all of $\boldsymbol{D}$ except any one row learns little from $\mathcal{P}(\boldsymbol{D})$ about the values in that unknown row when ϵ is small (Barrientos *et al.*, 2019).

The definition of ϵ-DP satisfies several desirable properties. Let $\mathcal{P}_1$ and $\mathcal{P}_2$ be ϵ_1-DP and ϵ_2-DP algorithms. The first is sequential composition: for any database $\boldsymbol{D}$, release of both $\mathcal{P}_1(\boldsymbol{D})$ and $\mathcal{P}_2(\boldsymbol{D})$ satisfies $(\epsilon_1 + \epsilon_2)$-DP. This means that we are able to calculate the total privacy leakage from releasing multiple statistics. The second is parallel composition. Let $\boldsymbol{D}_1$ and $\boldsymbol{D}_2$ be two data files on disjoint sets of individuals. Release of both $\mathcal{P}_1(\boldsymbol{D}_1)$ and $\mathcal{P}_2(\boldsymbol{D}_2)$ satisfies $\max\{\epsilon_1, \epsilon_2\}$-DP. The third is the post-processing property. For any algorithm $\mathcal{P}_2$, releasing $\mathcal{P}_2(\mathcal{P}_1(\boldsymbol{D}))$ for any $\boldsymbol{D}$ satisfies ϵ_1-DP. In other words, post-processing the output of an ϵ_1-DP algorithm does not imply any additional privacy loss.

A commonly used method to ensure ϵ-DP is the Laplace mechanism. Let $f(\boldsymbol{D})$ be a function that takes $\boldsymbol{D}$ as an input and outputs some statistic in $\mathbb{R}^k$. For example, f might sum the elements of one of the columns in $\boldsymbol{D}$. We define the global sensitivity $s(f, ||\cdot||) = \max_{\boldsymbol{D},\boldsymbol{D}',d(\boldsymbol{D},\boldsymbol{D}')=1} ||f(\boldsymbol{D}) - f(\boldsymbol{D}')||$, where $d(\boldsymbol{D}, \boldsymbol{D}') = 1$ implies that the two databases differ by only one row, and $||\cdot||$ is a norm specific to the context. The Laplace mechanism computes $LM(\boldsymbol{D}) = f(\boldsymbol{D}) + \boldsymbol{\kappa}$, where $\boldsymbol{\kappa}$ is a $k \times 1$ vector of independent draws from a Laplace distribution with density $g(x|\lambda) = (1/2\lambda)\exp(-|x|/\lambda)$, where $\lambda = s(f, ||\cdot||)/\epsilon$ (Dwork, 2006). In Section 3, we use the Laplace mechanism to construct causal estimators and associated 95% confidence intervals satisfying ϵ-DP.

As part of our developments in subsequent sections, we use the *subsample and aggregate algorithm* (Nissim *et al.*, 2007) to satisfy ϵ-DP. The algorithm consists of a sampling step and an aggregating step. In the sampling step, we partition the confidential data $\boldsymbol{D}$ into M disjoint subsets $\boldsymbol{D}_1, \ldots, \boldsymbol{D}_M$ and compute $f(\boldsymbol{D}_m)$ in each $\boldsymbol{D}_m$. In the aggregation step,

we compute the average $f(\boldsymbol{D}_1, \ldots, \boldsymbol{D}_M) = \sum_{m=1}^{M} f(\boldsymbol{D}_m)/M$. For many f, any single observation affects the output from at most one of the partitions, i.e., the one it is randomly assigned to. For such f, the global sensitivity of $f(\boldsymbol{D}_1, \ldots, \boldsymbol{D}_M)$ generally is $1/M$ times that of $f(\boldsymbol{D})$. Using this sensitivity, we apply the Laplace mechanism to $f(\boldsymbol{D}_1, \ldots, \boldsymbol{D}_M)$ to create the differentially private statistic. The reduced sensitivity decreases the variance of the noise distribution, which in turn offers potential for increased accuracy of released results.

3 Differentially Private Estimation of WATE

We now construct a differentially private estimator of τ and its associated 95% interval estimate for the three target populations reviewed in Table 1. Our general strategy is to (i) find expressions for global sensitivities of the point and variance estimators, (ii) use the subsample and aggregate algorithm with these global sensitivities to generate noisy versions of the point and variance estimates, and (iii) apply a Bayesian inferential procedure to turn these noisy quantities into an interval estimate for τ.

We begin by finding a global sensitivity of $\hat{\tau}$ for binary outcomes. Determining a sharp bound on the sensitivity is tricky, since changing any one data point can change not only the outcomes but the propensity score estimation and hence weights for all individuals in (3). Instead, we use the coarse bound shown in Lemma 3.1.

Lemma 3.1 *For $y_i(0), y_i(1) \in \{0, 1\}$, if $0 < e(\boldsymbol{x}_i) < 1$ for all $i = 1, \ldots, n$, the global sensitivity of $\hat{\tau}$ in (3) for the ATE, ATT and ATC is bounded by 2, i.e., $s(\hat{\tau}, |\cdot|) \leq 2$.*

Proof For any $\boldsymbol{D}$, we have

$$
\begin{aligned}
|\hat{\tau}(\boldsymbol{D})| &\leq \left| \frac{\sum_{i=1}^{n} w_{1i} z_i y_i(1)}{\sum_{i=1}^{n} w_{1i} z_i} - \frac{\sum_{i=1}^{n} w_{0i}(1 - z_i) y_i(0)}{\sum_{i=1}^{n} w_{0i}(1 - z_i)} \right| \\
&\leq \max\{\max_{i=1:n} y_i(1) - \min_{i=1:n} y_i(0), \max_{i=1:n} y_i(0) - \min_{i=1:n} y_i(1)\} \leq 1, \qquad (7)
\end{aligned}
$$

since $y_i(0), y_i(1) \in \{0, 1\}$, $w_{0i} > 0$ for at least one $z_i = 0$, and $w_{1i} > 0$ for at least one $z_i = 1$. Thus, for any two neighboring datasets $\boldsymbol{D}$ and $\boldsymbol{D}'$, we have $s(\hat{\tau}, |\cdot|) \leq |\hat{\tau}(\boldsymbol{D})| + |\hat{\tau}(\boldsymbol{D}')| \leq 2$.

To ensure all $0 < w_{0i}, w_{1i} < \infty$ for $i = 1, \ldots, n$, we use truncated propensity scores for

the differentially private WATE. The truncation limit a is set before looking at values in $\boldsymbol{D}$, so that a is not data-dependent. Under truncation, the global sensitivities for the estimated treatment effects remain bounded by two. The truncation also facilitates bounding the sensitivities for the estimated variances, as we now show.

Theorem 3.2 *For $y_i(0), y_i(1) \in \{0,1\}$, bounds on the global sensitivities for $\hat{V}^T_{ATE}$, $\hat{V}^T_{ATT}$ and $\hat{V}^T_{ATC}$ are as follows: $s(\hat{V}^T_{ATE}, |\cdot|) \leq \frac{1}{an}$; $s(\hat{V}^T_{ATT}, |\cdot|) \leq \frac{1}{2a^2n}$; and $s(\hat{V}^T_{ATC}, |\cdot|) \leq \frac{1}{2a^2n}$.*

Proof For any $\boldsymbol{D}$, we have

$$\hat{V}^T_{ATE}(\boldsymbol{D}) = \frac{\sum_{i=1}^n \left[\frac{v_1(\boldsymbol{x}_i)}{e^T(\boldsymbol{x}_i)} + \frac{v_0(\boldsymbol{x}_i)}{1-e^T(\boldsymbol{x}_i)}\right]}{n^2} \leq \frac{n(\frac{2}{4a})}{n^2} = \frac{1}{2an}, \tag{8}$$

where the inequality follows because $v_0(\boldsymbol{x}_i), v_1(\boldsymbol{x}_i) \leq 1/4$ and $\max\{e^T(\boldsymbol{x}_i), 1-e^T(\boldsymbol{x}_i)\} \geq a$. Hence, for any two neighboring $\boldsymbol{D}$ and $\boldsymbol{D}'$, we have $s(\hat{V}^T_{ATE}, |\cdot|) \leq |\hat{V}^T_{ATE}(\boldsymbol{D})| + |\hat{V}^T_{ATE}(\boldsymbol{D}')| \leq \frac{1}{an}$. Applying similar logic for $\hat{V}^T_{ATT}(\boldsymbol{D})$, we have

$$\hat{V}^T_{ATT}(\boldsymbol{D}) = \frac{\sum_{i=1}^n e^T(\boldsymbol{x}_i)^2 \left[\frac{v_1(\boldsymbol{x}_i)}{e^T(\boldsymbol{x}_i)} + \frac{v_0(\boldsymbol{x}_i)}{1-e^T(\boldsymbol{x}_i)}\right]}{\left\{\sum_{i=1}^n e^T(\boldsymbol{x}_i)\right\}^2} \leq \frac{\sum e^T(\boldsymbol{x}_i)\left[1/4 + \frac{(1-a)}{4a}\right]}{\left\{\sum_{i=1}^n e^T(\boldsymbol{x}_i)\right\}^2} = \frac{\left[1/4 + \frac{(1-a)}{4a}\right]}{\left[\sum_{i=1}^n e^T(\boldsymbol{x}_i)\right]} \leq \frac{1}{4na^2}. \tag{9}$$

Hence, $s(\hat{V}^T_{ATT}, |\cdot|) \leq |\hat{V}^T_{ATT}(\boldsymbol{D})| + |\hat{V}^T_{ATT}(\boldsymbol{D}')| \leq \frac{1}{2na^2}$. Likewise, for $\hat{V}^T_{ATC}(\boldsymbol{D})$, we have

$$\begin{aligned}\hat{V}^T_{ATC}(\boldsymbol{D}) &= \frac{\sum_{i=1}^n (1-e^T(\boldsymbol{x}_i))^2 \left[\frac{v_1(\boldsymbol{x}_i)}{e^T(\boldsymbol{x}_i)} + \frac{v_0(\boldsymbol{x}_i)}{1-e^T(\boldsymbol{x}_i)}\right]}{\left[\sum_{i=1}^n (1-e^T(\boldsymbol{x}_i))\right]^2} = \frac{\sum_{i=1}^n (1-e^T(\boldsymbol{x}_i))\left[\frac{v_1(\boldsymbol{x}_i)(1-e^T(\boldsymbol{x}_i))}{e^T(\boldsymbol{x}_i)} + v_0(\boldsymbol{x}_i)\right]}{\left[\sum_{i=1}^n (1-e^T(\boldsymbol{x}_i))\right]^2} \\ &\leq \frac{\sum_{i=1}^n (1-e^T(\boldsymbol{x}_i))\left[1/4 + \frac{(1-a)}{4a}\right]}{\left[\sum_{i=1}^n (1-e^T(\boldsymbol{x}_i))\right]^2} = \frac{\left[1/4 + \frac{1-a}{4a}\right]}{\left[\sum_{i=1}^n (1-e^T(\boldsymbol{x}_i))\right]} \leq \frac{1}{4na^2}. \end{aligned} \tag{10}$$

Hence, $s(\hat{V}^T_{ATC}, |\cdot|) \leq |\hat{V}^T_{ATC}(\boldsymbol{D})| + |\hat{V}^T_{ATC}(\boldsymbol{D}')| \leq \frac{1}{2na^2}$.

To avoid introducing a substantial bias in $\hat{\tau}$, we should make a small, e.g., $a \leq 0.05$. However, with small a, the global sensitivities in Theorem 3.2 could be large enough that, with small ϵ, the noise variance in the Laplace mechanism is large compared to $\hat{V}$ itself, which could lead to undesirably wide confidence intervals.

Therefore, rather than use Laplace mechanisms for $\hat{\tau}$ and $\hat{V}$, we use the subsampling and aggregation technique reviewed in Section 2.2. We split $\boldsymbol{D}$ into M disjoint subsets, $\{\boldsymbol{D}_1, \ldots, \boldsymbol{D}_M\}$, of approximately equal size. In each $\boldsymbol{D}_m$, we estimate propensity scores using only the data in that subset, and truncate them as in (6). Using the truncated propensity scores, in each $\boldsymbol{D}_m$ we compute the treatment effect estimate $\hat{\tau}_m^T$ of interest and its approximated variance $\hat{V}_m^T$ using the expressions in Table 1, replacing each $e(\boldsymbol{x}_i)$ with its truncated version $e^T(\boldsymbol{x}_i)$. Finally, we average these estimates over the M partitions to obtain, for the particular treatment effect estimate of interest,

$$\bar{\tau}^T = \sum_{m=1}^{M} \hat{\tau}_m^T / M, \;\; \bar{V}^T = \sum_{m=1}^{M} \hat{V}_m^T / M. \tag{11}$$

The global sensitivities of $\bar{\tau}^T$ and $\bar{V}^T$ are $2/M$ and $s(\hat{V}^T, |\cdot|)/M$, respectively.

To complete the subsampling and aggregation algorithm, we add independent noise to each of the quantities in (11) using Laplace mechanisms. Suppose the total privacy budget is ϵ. For $0 < \pi < 1$, we use $(1-\pi)\epsilon$ privacy budget for the treatment effect estimate and $\pi\epsilon$ for the variance estimate. Specifically, we compute $\bar{\tau}^{T,\epsilon} = \bar{\tau}^T + \eta_1$, where $\eta_1 \sim \mathit{Laplace}(0, 2/(M\epsilon(1-\pi)))$ and $\bar{V}^{T,\epsilon} = \bar{V}^T + \eta_2$, where $\eta_2 \sim \mathit{Laplace}(0, s(\bar{V}^T, |\cdot|)/(M\epsilon\pi))$.

While $\bar{\tau}^{T,\epsilon}$ and $\bar{V}^{T,\epsilon}$ are differentially private, they may not be readily usable to make interpretable inferences about τ. In particular, $\bar{\tau}^{T,\epsilon}$ is not guaranteed to lie in $(-1, 1)$, and $\bar{V}^{T,\epsilon}$ could be negative. Therefore, we use a Bayesian post-processing algorithm to turn the differentially private point and variance estimates into interpretable inferences about τ. The basic idea is as follows. Since the data analyst only has $(\bar{\tau}^{T,\epsilon}, \bar{V}^{T,\epsilon})$, the analyst treats $(\bar{\tau}^T, \bar{V}^T)$ as unknown quantities. The analyst draws many, say L, plausible values of the unobserved $(\bar{\tau}^T, \bar{V}^T)$ from their posterior distribution, and from each plausible value samples a value of τ. The L draws of τ can be summarized for inferences.

We now offer the details of this post-processing step. For clarity we introduce new notation $(\bar{\tau}, \bar{V})$ to represent the analyst's random variables for the unknown values of $(\bar{\tau}^T, \bar{V}^T)$.

We specify two models $\mathcal{M}_1$ and $\mathcal{M}_2$ independently, given by

$$\begin{aligned}
&(\text{Model } \mathcal{M}_1): \quad \bar{\tau}^{T,\epsilon} = \bar{\tau} + \zeta_2, \;\; \zeta_2 \sim \text{Laplace}(0, 2/(M\epsilon(1-\pi))), \\
&(\text{Model } \mathcal{M}_2): \quad \bar{V}^{T,\epsilon} = \bar{V} + \zeta_1, \;\; \zeta_1 \sim \text{Laplace}(0, s(\bar{V}^T, |\cdot|)/(M\epsilon\pi)).
\end{aligned} \tag{12}$$

We assign $\bar{\tau} \sim U(-1, 1)$ and $\bar{V} \sim U(0, s(\bar{V}^T, |\cdot|)/2)$ prior distributions, where $s(\bar{V}^T, |\cdot|)/2$ is the upper bound from Theorem 3.2 for the treatment effect of interest. We estimate the posterior distributions of $\bar{\tau}$ and $\bar{V}$ using elliptical slice sampling (Nishihara *et al.*, 2014). Importantly, we do not use the confidential values of $\bar{\tau}^T$ and $\bar{V}^T$ in the sampling algorithms; we only use the differentially private statistics.

We obtain L post burn-in samples of $\bar{\tau}$, denoted as $\bar{\tau}^{*(1)}, \ldots, \bar{\tau}^{*(L)}$, and of $\bar{V}$, denoted as $\bar{V}^{*(1)}, \ldots, \bar{V}^{*(L)}$. For $l = 1, \ldots, L$, we draw a sample $\tilde{\tau}^{(l)} \sim N(\bar{\tau}^{*(l)}, \bar{V}^{*(l)})$. These L draws represent an approximate posterior distribution for τ. We use $\tilde{\tau}^{\epsilon} = \sum_{l=1}^{L} \tilde{\tau}^{(l)}/L$ as the privacy-protected point estimator for τ. We construct the 2.5% and 97.5% empirical quantiles $\tilde{\tau}^{\epsilon}_{lower}$ and $\tilde{\tau}^{\epsilon}_{upper}$ from $\tilde{\tau}^{(1)}, \ldots, \tilde{\tau}^{(L)}$, and use $(\tilde{\tau}^{\epsilon}_{lower}, \tilde{\tau}^{\epsilon}_{upper})$ as the privacy-protected 95% interval estimate for τ.

In drawing values of τ, we rely on large-sample normality for the sampling distribution of $\bar{\tau}^T$ as defined in (11), namely $\bar{\tau}^T \sim N(\tau, \bar{V}^T)$. With a diffuse prior on τ, we have $\tau \sim N(\bar{\tau}^T, \bar{V}^T)$, which we can sample from to summarize inferences about τ. This presumes the sampling variability in $\bar{V}^T$ is negligible compared to $\bar{V}^T$ itself, which is generally reasonable and typically assumed in large-sample inference (Rubin, 1987). In our setting, the analyst does not know $\bar{\tau}^T$ and $\bar{V}^T$; rather, the analyst has plausible draws of each. Thus, we follow the strategy described in Zhou and Reiter (2010) for Bayesian inference with plausible draws of unknown values: for each plausible draw $(\tilde{\tau}^{*(l)}, \bar{V}^{*(l)})$ of $(\bar{\tau}^T, \bar{V}^T)$, we sample a value of τ from $N(\bar{\tau}^{*(l)}, \bar{V}^{*(l)})$, and concatenate the draws for inferences about τ.

The entire process for estimating τ is summarized in Algorithm 1. Theorem 3.3 formally states and proves that Algorithm 1 is differentially private.

Theorem 3.3 *Algorithm 1 satisfies $\epsilon-$differential privacy.*

Proof Since $\bar{\tau}^T$ has a global sensitivity of $2/M$, defining $\bar{\tau}^{T,\epsilon} = \bar{\tau}^T + Laplace(0, 2/(M\epsilon(1-\pi)))$ is a $(1-\pi)\epsilon$-DP algorithm. Using a similar argument, $\bar{V}^{T,\epsilon} = \bar{V}^T + Laplace(0, s(\bar{V}^T, |\cdot$

|)/($M\epsilon\pi$)) is a $\pi\epsilon$-DP algorithm. The Bayesian inference steps rely entirely on $(\bar{\tau}^{T,\epsilon}, \bar{V}^{T,\epsilon})$. By the post-processing property of differential privacy, they do not affect the privacy guarantee. Hence, releasing $\tilde{\tau}^{\epsilon}$ and $(\tilde{\tau}^{\epsilon}_{lower}, \tilde{\tau}^{\epsilon}_{upper})$ from Algorithm 1 is $(1-\pi)\epsilon + \pi\epsilon = \epsilon$-DP.

3.1 Theoretical Study of the DP WATE

In this section, we discuss some asymptotic properties of $\tilde{\tau}^{\epsilon}$. Here, we presume any bias in the treatment effect estimators introduced by truncating propensity scores is negligible. This is reasonable when a is small and all $e(\boldsymbol{x}_i)$ are bounded away from a. If the truncation does produce a non-negligible bias, the results still support the statement that the differentially private WATE has a similar distribution as the WATE based on the trimmed propensity scores.

As discussed in Section 3, when the sample size n is large, the analyst's draws from $N(\bar{\tau}, \bar{V})$ are samples from the posterior distribution of τ, from which the analyst obtains $\tilde{\tau}^{\epsilon}$. Let $\tilde{g}^{\epsilon}$ denote this distribution, i.e., $\tilde{g}^{\epsilon} = N(\bar{\tau}, \bar{V})$. When n is large, the distribution of the non-private estimator $\hat{\tau}$ is approximately $g = \mathrm{N}(\tau, V)$. To assess the discrepancy between $\tilde{\tau}^{\epsilon}$ and $\hat{\tau}$, we develop an upper bound for the discrepancy between g and $\tilde{g}^{\epsilon}$, given by $P(KL(\tilde{g}^{\epsilon}, g) > c)$, for any $c > 0$, as a function of M, ϵ and V, as $n \to \infty$. Here, KL stands for Kullback-Leibler divergence.

For $m = 1, \ldots, M$, let n_m be the number of observations in $\boldsymbol{D}_m$; let $\boldsymbol{D}_{T,m}$, $\boldsymbol{D}_{u,m}$ and $\boldsymbol{D}_{l,m}$ denote the sets of observations with propensity scores between a and $1-a$, greater than $1-a$, and less than a, respectively; and, let $n_{T,m}$, $n_{u,m}$ and $n_{l,m}$ be the number of samples in each of $\boldsymbol{D}_{T,m}$, $\boldsymbol{D}_{u,m}$ and $\boldsymbol{D}_{l,m}$, respectively. We make the following assumptions about the sample sizes.

(A1) As $n \to \infty$, $n_m \to \infty$ for all $m = 1, \ldots, M$.

(A2) As $n_m \to \infty$, $n_{u,m}/n_m \to 0$ and $n_{l,m}/n_m \to 0$.

Since $n_m = n_{T,m} + n_{l,m} + n_{u,m}$, assumptions (A1) and (A2) imply that $n_{T,m}/n_m \to 1$ as $n \to \infty$.

Lemma 3.4 *Under (A1) and (A2), the following results hold.*

(i) $P(|\bar{\tau} - \tau| > c) \leq 2\exp(-M\epsilon(1-\pi)c/6)$, *as* $n \to \infty$, *for any* $c > 0$.

Algorithm 1: Differentially Private WATE and its 95% Interval Estimate

Input: (1) $\boldsymbol{D}$: Dataset $\{y_i, \boldsymbol{x}_i, z_i : i = 1, \ldots, n\}$; (2) M : Number of partitions; (3) a : Truncation level; (4) ϵ : Privacy loss budget; (5) π : fraction of privacy loss budget allocated to variance estimation.

Output: (1) DP WATE estimate $\tilde{\tau}^{\epsilon}$; (2) DP 95% interval $(\tilde{\tau}^{\epsilon}_{lower}, \tilde{\tau}^{\epsilon}_{upper})$ for WATE

begin

`/* Step 1: Partition the data as a part of subsample and aggregation step. */`

Choose a random partition $\{\boldsymbol{D}_1, \ldots, \boldsymbol{D}_M\}$ of $\boldsymbol{D}$

`/* Step 2: Compute WATE estimate and its estimated variance based on truncated propensity scores in each subset. */`

for $m \in 1 : M$ **do**

Compute WATE estimate $\hat{\tau}^T_m$ and its approximated variance $\hat{V}^T_m$ using the truncated propensity score defined in (6) from $\boldsymbol{D}_m$.

end

`/* Step 3: Add noise following Laplace mechanism. */`

Compute the average of treatment effects $\bar{\tau}^T$ and its estimated average variance $\bar{V}^T$ following (11).

Generate $\eta_1 \sim \text{Laplace}(0, 2/(M\epsilon(1-\pi)))$ and $\eta_2 \sim \text{Laplace}(0, s(\bar{V}^T, |\cdot|)/(\epsilon\pi))$.

Compute noisy versions $\bar{\tau}^{T,\epsilon} = \bar{\tau}^T + \eta_1$ and $\bar{V}^{T,\epsilon} = \bar{V}^T + \eta_2$.

`/* Step 4: Apply Bayesian post-processing steps. */`

Fit models $\mathcal{M}_1$ and $\mathcal{M}_2$ independently, which are given by

$$\begin{aligned} &(\text{Model } \mathcal{M}_1): \quad \bar{\tau}^{T,\epsilon} = \bar{\tau} + \zeta_2, \ \ \zeta_2 \sim \text{Laplace}(0, 2/(M\epsilon(1-\pi))), \\ &(\text{Model } \mathcal{M}_2): \quad \bar{V}^{T,\epsilon} = \bar{V} + \zeta_1, \ \ \zeta_1 \sim \text{Laplace}(0, s(\bar{V}^T, |\cdot|)/(M\epsilon\pi)), \end{aligned} \tag{13}$$

for $l \in 1 : L$ **do**

Draw elliptical slice samples (Nishihara *et al.*, 2014) for $\bar{\tau}$ and $\bar{V}$, denoted by $\bar{\tau}^{(l)}$ and $\bar{V}^{(l)}$, respectively.

Draw $\tilde{\tau}^{(l)} \sim N(\bar{\tau}^{(l)}, \bar{V}^{(l)})$.

end

Compute $\tilde{\tau}^{\epsilon} = \sum_{l=1}^{L} \tilde{\tau}^{(l)}/L$ and $\tilde{\tau}^{\epsilon}_{lower} = 2.5\%$ empirical quantile, $\tilde{\tau}^{\epsilon}_{upper} = 97.5\%$ empirical quantile.

return $\tilde{\tau}^{\epsilon}$, $(\tilde{\tau}^{\epsilon}_{lower}, \tilde{\tau}^{\epsilon}_{upper})$.

end

(ii) $P(|\bar{V} - V| > c) \leq 2\exp(-M\epsilon\pi c/6)$, *as* $n \to \infty$, *for any* $c > 0$.

A proof of Lemma 3.4 is in the appendix. We use Lemma 3.4 to derive a bound on $P(KL(\tilde{g}^{\epsilon}, g) > c)$.

Theorem 3.5 *Under (A1) and (A2), we have*

$$P(KL(\tilde{g}^{\epsilon}, g) > c) \leq 2\exp\left(-\frac{M\epsilon(1-\pi)\sqrt{2Vc}}{6\sqrt{3}}\right) + 4\exp(-M\epsilon\pi Vc/9),$$

as $n \to \infty$.

Theorem (3.5) shows that, as M or ϵ get large, the distance between the two quantities goes to 0.

4 Simulation Studies

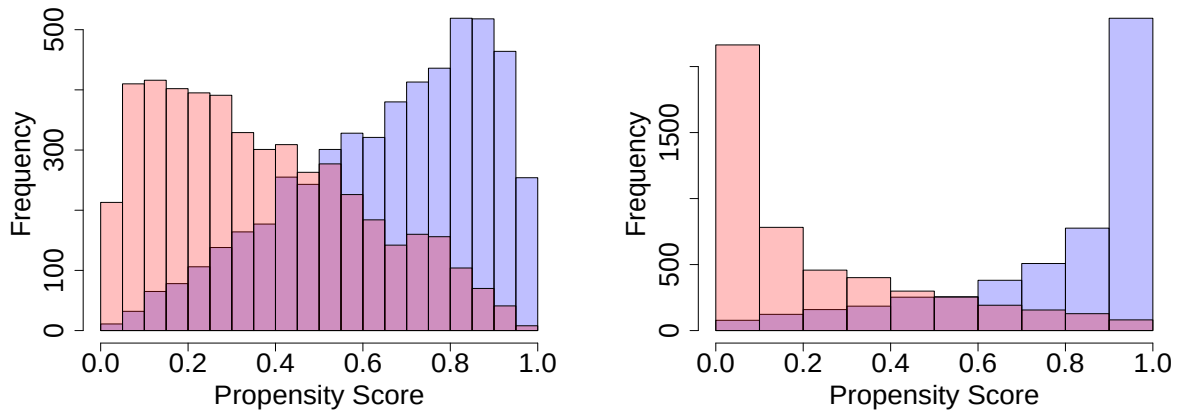
In this section, we illustrate repeated sampling properties of the differentially private point, variance, and interval estimates of the ATE, ATT and ATC. We begin with simulations that set $M = 100$ and $a = 0.05$ for data of size $n = 10000$ and privacy loss budget $\epsilon = 1$. We then vary simulation design parameters one at a time to investigate the sensitivity of the findings. Section 4.2 varies M; Section 4.3 varies a; Section 4.4 varies n; and, Section 4.5 varies ϵ.

4.1 Baseline Studies with $M = 100, a = 0.05, n = 10000, \epsilon = 1$

To create $\boldsymbol{D}$ in any simulation run, we generate $n = 10000$ observations each measured on $p = 4$ covariates, $\boldsymbol{x}_i = (x_{i1}, x_{i2}, x_{i3}, x_{i4})$. We simulate $\boldsymbol{x}_i \sim N(\boldsymbol{0}, (1-\rho)\boldsymbol{I} + \rho\boldsymbol{J})$, $0 \leq \rho \leq 1$, where $\boldsymbol{J}$ is a $p \times p$ matrix with each entry as 1. The covariance implies correlation of ρ between any two predictors, and we set $\rho = 0.2$ in all simulations. For $i = 1, \ldots, n$, we generate a treatment status z_i from a Bernoulli draw with probability $P(z_i = 1|\boldsymbol{x}_i)$, where

$$\text{logit}[P(z_i = 1|\boldsymbol{x}_i)] = 0.1 + 0.2\eta x_{i1} + 0.5\eta x_{i2} - 0.25\eta x_{i3} - 0.45\eta x_{i4}. \tag{14}$$

We vary $\eta \in \{2, 4\}$ to change the level of overlap in the treatment and control samples, as shown in Figure 1. As η increases, we observe increasingly sparse overlap in the propensity



(a) $\eta = 2$ (b) $\eta = 4$

Figure 1: Simulated propensity score distributions in the simulation of Section 4.1 for treated (purple) and controls (pink). The propensity score distributions are shown for $\eta = 2$ and $\eta = 4$. The propensity score distributions show much less overlap when $\eta = 4$.

score distribution.

For each $(\boldsymbol{x}_i, z_i)$, we simulate the potential binary outcomes, $(y_i(0), y_i(1))$, from Bernoulli distributions with probabilities governed by

$$\text{logit}[P(y(z) = 1)] = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \gamma z. \tag{15}$$

We set $(\beta_0, \beta_1, \beta_2, \beta_3, \beta_4) = (0.15, -0.2, 0.3, -0.4, 0.6)$. We also vary $\gamma \in \{0, 1, 2\}$ to represent different strengths of treatment effects. Thus, we have six simulation scenarios corresponding to each combination of $(\eta, \gamma) \in \{2, 4\} \times \{0, 1, 2\}$.

In each simulation, we compute the true treatment effects,

$$\tau_{ATE} = (1/n) \sum_{i=1}^{n} \{P(y_i(1) = 1 \,|\, \boldsymbol{x}_i) - P(y_i(0) = 1 \,|\, \boldsymbol{x}_i)\} \tag{16}$$

$$\tau_{ATT} = (1/n_T) \sum_{i:z_i=1} \{P(y_i(1) = 1 \,|\, \boldsymbol{x}_i) - P(y_i(0) = 1 \,|\, \boldsymbol{x}_i)\} \tag{17}$$

$$\tau_{ATC} = (1/n_C) \sum_{i:z_i=0} \{P(y_i(1) = 1 \,|\, \boldsymbol{x}_i) - P(y_i(0) = 1 \,|\, \boldsymbol{x}_i)\}, \tag{18}$$

where the expressions for $P(y_i(1) = 1 \,|\, \boldsymbol{x}_i)$ and $P(y_i(0) = 1 \,|\, \boldsymbol{x}_i)$ are obtained from (15). The

quantities n_T and n_C are the number of treated and control subjects, respectively.

For comparisons, in each simulation run, we compute estimated treatment effects $\hat{\tau}$ for each causal estimand without privacy concerns, i.e., without the partitions, truncation, or Laplace noise. We also compute the 95% confidence interval (CI) using $(\hat{\tau} \pm 1.96\sqrt{\hat{V}})$, where $\hat{V}$ is the estimated variance of the treatment effect calculated on the sample using (5) without any privacy protections. We run 500 independent replications of each scenario, sampling a new set of $\{(\boldsymbol{x}_1, z_1, y_1), \ldots, (\boldsymbol{x}_n, z_n, y_n)\}$ each time. We equally allocate privacy budget in the WATE point and variance estimation, and set $\pi = 0.5$.

As evident in Table 2, the differentially private point estimates have small average errors, indicating that they accurately estimate true treatment effects. For context, the values of the various τ for scenarios with $\gamma = 1, 2$ tend to be around .20 to .30, so that RMSEs of .01 to .02 are modest fractions of the true treatment effects. The average errors from the differentially private estimates tend to be larger than those from the non-privacy protected point estimates, reflecting the combined effects of the Laplace noise, truncation limits, and subsampling. All methods are more accurate when there is greater overlap in the propensity scores, as one would expect. The 95% CIs for the various τ without privacy protection cover less often than the nominal 95% rate, whereas the 95% intervals based on the differentially private algorithms tend to cover more often than the nominal 95% rate. The steps taken to protect privacy result in increased average interval lengths, which seemingly is the price to pay for the privacy protection.

4.2 Sensitivity to the Choice of M

The choice of $M = 100$ ensures that the standard deviation of the Laplace noise is much smaller than $\bar{V}^T$. In this section, we consider the effects on inference of changing M by considering $M \in \{50, 100, 200\}$. In generating the data, we use the more challenging case of less overlap between treatment and control propensity score distributions, setting $\eta = 4$ and $\gamma = 1$. All other parameters are set at the values described in Section 4.1.

Figure 2 displays the RMSEs of the point estimators, and the coverage rates and average lengths of the 95% intervals. In these simulations, we see little practical impact of changing M on the properties of the differentially private WATE estimates. For all M considered, the

differentially private WATE estimates generally are close to the corresponding non-private treatment effect WATE estimates computed with $\boldsymbol{D}$, and the coverage rates of the differentially private intervals are all around 98%. The interval lengths also are unremarkably different, with some suggestion that the interval lengths are smallest at $M = 50$. As M increases, the intervals are subject to two countervailing effects. The variance of the Laplace noise decreases thereby encouraging shorter intervals, and the uncertainty in the propensity score estimates in each partition increases thereby encouraging longer intervals. We confirmed the latter fact by constructing 95% intervals from samples of $\tau \sim N(\bar{\tau}^T, \bar{V}^T)$, that is, using partitioning without adding Laplace noise. We suggest a rule-of-thumb for selecting M in Section 6.

4.3 Sensitivity to the Choice of Truncation Point

While $a = .05$ may be a reasonable cut-off for propensity scores in many contexts, it is instructive to investigate the performance of the differentially private WATE inferences for other realistic values of a. To this end, we also examine the differentially private WATE inferences for $a \in \{0.03, 0.07, 0.1\}$. Let $\tilde{\tau}^{\epsilon,a}$ correspond to the differentially private point estimate for the WATE for truncation limit a. We present the absolute distance between the non-private estimator on $\hat{\tau}$ computed using $\boldsymbol{D}$, i.e., without truncation or privacy protections, and $\tilde{\tau}^{\epsilon,a}$. We write this difference as $\text{Dev}(\tilde{\tau}^{\epsilon,a}) = |\tilde{\tau}^{\epsilon,a} - \hat{\tau}|$. We also present coverage rates and average lengths for the 95% intervals using different choices of a. We set $\epsilon = 1$, $M = 100$, $\eta = 4$, and $\gamma = 1$.

As evident in Table 3, in these simulations the properties of the point estimate and coverage rates for ATE, ATT and ATC are similar for all values of a investigated. This is because the Laplace noise variances are of comparable magnitude for the values of a in this range. However, if we decrease a to values near zero, say $a = .001$, the variance in the Laplace mechanism applied to $\bar{V}^T$ is about 20 times higher compared to using $a = 0.05$. As a result, the 95% intervals are much wider, and their coverage rate approaches 1. When $a = 0.001$, the RMSE values are nearly 1.5 times higher than when $a = 0.05$.

4.4 Sensitivity to the Sample Size

We next illustrate the effect of the sample size by repeating simulations with $n = 100000$ and $n = 5000$. We set $\epsilon = 1$, $M = 100$, $a = 0.05$, $\eta = 4$, and $\gamma = 1$ for this simulation. Table 4 displays the results. The point estimates from the differentially private WATE estimators remain accurate, with RMSE values decreasing as n increases. When $n = 5000$, the interval estimates widen substantially for both the non-private and differentially private estimators. At $n = 5000$, the coverage rate for the differentially private interval is near 100%, indicating that the inferential procedure at this sample size results in overly wide intervals.

4.5 Sensitivity to the value of ϵ

Finally, we consider the effects of changing ϵ. In general, the choice of ϵ is driven by privacy desiderata, for example, as specified by the data holders. Ideally, the value of ϵ also takes into account the likely usefulness of the outputs that could result from applying the differentially private algorithm; that is, the process of setting ϵ considers a trade off between risk and usefulness. Here, we examine results for $\epsilon \in \{0.5, 1, 5\}$. As in Section 4.2, to generate $\boldsymbol{D}$ in each of the 500 simulation runs, we set $\eta = 4$ and $\gamma = 1$.

Figure 3 displays the results. Changing ϵ in these simulations influences the estimated variances and hence lengths of the interval estimates. For small values of ϵ, the differentially private 95% intervals are wider and have larger coverage rates. As ϵ increases, the coverage rates become closer to nominal.

5 Illustration with the Adult Income Data

We demonstrate the application of the differentially private WATE estimators using the "Adult" data set (Becker and Kohavi, 1996), which we accessed via the UCI Machine Learning Repository (`https://archive.ics.uci.edu/`). We emphasize that this example serves to illustrate the methodology and is not intended to be a thorough causal analysis of the effect of education on income.

The data comprise $n = 30162$ individuals with complete information on the following variables: the age of the individual; the marital status, which includes seven categories - married to a civilian spouse, married to a spouse in the armed forces, married but with an

absentee spouse, never married, divorced, separated, or widowed; the race, which includes five categories - white, black, American Indian Eskimo, Asian Pacific islander, and other; the sex, which includes two categories - male and female; the individual's occupation, which spans across 15 categories including executive-managerial, farming, fishing, transportation, sales, administrative-clerical, and more; and, an indicator of whether their native country is the USA or not. We classify individuals as treated ($z = 1$) if they have earned a bachelor's degree or higher, and as controls ($z = 0$) if they have an education level lower than a bachelor's degree. We make a binary outcome from income as below \$50000 ($y = 0$) or at least \$50000 ($y = 1$).

We estimate propensity scores using a logistic regression of z on age, marital status, race, sex, occupation, and a binary indicator denoting whether the individual's native country is the USA. We use Algorithm 1 to make differentially private inferences about the effect of education on income. Given the relatively large sample size, for this analysis, we set $M = 100$. This M value is sufficiently large enough that the Laplace noise variance is expected to be less than the within-partition variance of the WATE. We set $a = 0.05$ and $\pi = 0.5$.

Table 5 presents results for two values of $\epsilon \in \{0.5, 1\}$. For both the differentially private WATE estimates and their non-privacy protected counterparts, the point estimates are positive and 95% confidence intervals exclude zero. These suggest a positive association between years of education and income level. The differentially private WATE estimates and their non-privacy protected analogues are close to one another. The 95% intervals constructed from Algorithm 1 are wider than the corresponding intervals based on the full data.

6 Concluding Remarks

We present an approach to estimating weighted average treatment effects with binary outcomes while ensuring differential privacy. Simulation and empirical results suggest that the approach can result in accurate point estimates with conservative interval coverage rates. To implement the approach, analysts need to select M and a. We now suggest some rule-of-thumb guidance for these choices. We note that analysts can undertake simulation studies akin to those presented here to facilitate choices tuned more closely to their settings.

In determining appropriate values for M, we recommend analysts to opt for the smallest feasible M while ensuring that, given the privacy parameter ϵ, the standard deviation of the Laplace noise from the subsampling and aggregation process remains significantly smaller than the sensitivity of the estimated variance of the WATE estimate obtained from the full data. When such an M cannot be found, the noise from privacy protection can overshadow the noise from sampling variability, thereby compromising the reliability of the ensuing estimates.

In general, we recommend selecting a to be sufficiently small to minimize the likelihood of truncating propensity scores, yet not so small as to inflate the sensitivity of the variance estimators. In essence, a can be adjusted to ensure that the sensitivity of the variance estimator is acceptable, particularly for the values of M under consideration.

7 Appendix

This section presents proofs of the theoretical results outlined in Section 3.1.

7.1 Proof of Lemma 3.4

We begin by proving part (i) of the lemma. Note that,

$$\begin{aligned} P(|\bar{\tau} - \tau| > c) \leq\ & P(|\bar{\tau} - \bar{\tau}^{T,\epsilon}| > c/3) \\ & + P(|\bar{\tau}^{T} - \bar{\tau}^{T,\epsilon}| > c/3) + P(|\bar{\tau}^{T} - \tau| > c/3). \end{aligned} \tag{19}$$

The first and second terms correspond to probabilities under the Laplace distribution. More specifically,

$$\begin{aligned} &P(|\bar{\tau}^{T} - \bar{\tau}^{T,\epsilon}| > c/3) = E_{y,\boldsymbol{x},z} P(|\bar{\tau}^{T} - \bar{\tau}^{T,\epsilon}| > c/3|\boldsymbol{D}) = \exp(-M\epsilon(1-\pi)c/6) \\ &P(|\bar{\tau} - \bar{\tau}^{T,\epsilon}| > c/3) = E_{y,\boldsymbol{x},z} P(|\bar{\tau} - \bar{\tau}^{T,\epsilon}| > c/3|\boldsymbol{D}) \leq \exp(-M\epsilon(1-\pi)c/6). \end{aligned} \tag{20}$$

It remains to show the bound for $P(|\bar{\tau}^T - \tau| > c/3)$. To this end, note that

$$(c^2/9)P(|\bar{\tau}^T - \tau| > c/3) \leq E_{y,\boldsymbol{x},z}[(\bar{\tau}^T - \tau)^2] \leq 2Var(\bar{\tau}^T) + 2\{E_{y,\boldsymbol{x},z}[\bar{\tau}^T] - \tau\}^2$$
$$= (2/M^2)\sum_{m=1}^{M} Var(\hat{\tau}_m^T) + 2\{(1/M)\sum_{m=1}^{M} E_{y,\boldsymbol{x},z}[\hat{\tau}_m^T] - \tau\}^2. \tag{21}$$

Let $\boldsymbol{D}_{u,m}, \boldsymbol{D}_{l,m}, \boldsymbol{D}_{T,m}$ denote the samples within the mth partition $\boldsymbol{D}_m$ with propensity score greater than $1-a$, less than a and between a and $(1-a)$, respectively. Now observe that

$$\frac{1}{n_m}\sum_{i=1}^{n_m} w_{1i} z_i y_i(1) = \frac{1}{n_m}\sum_{i\in\boldsymbol{D}_{T,m}} w_{1i} z_i y_i(1) + \frac{1}{n_m}\sum_{i\in\boldsymbol{D}_{l,m}} w_{1i} z_i y_i(1) + \frac{1}{n_m}\sum_{i\in\boldsymbol{D}_{u,m}} w_{1i} z_i y_i(1)$$
$$= \frac{1}{n_m}\sum_{i\in\boldsymbol{D}_{T,m}} \frac{t(\boldsymbol{x}_i)}{e^T(\boldsymbol{x}_i)} z_i y_i(1) + \frac{1}{n_m}\sum_{i\in\boldsymbol{D}_{l,m}} \frac{t(\boldsymbol{x}_i)}{e^T(\boldsymbol{x}_i)} z_i y_i(1) + \frac{1}{n_m}\sum_{i\in\boldsymbol{D}_{u,m}} \frac{t(\boldsymbol{x}_i)}{e^T(\boldsymbol{x}_i)} z_i y_i(1)$$
$$= \frac{1}{n_m}\sum_{i\in\boldsymbol{D}_{T,m}} \frac{t(\boldsymbol{x}_i)}{e(\boldsymbol{x}_i)} z_i y_i(1) + \frac{1}{n_m a}\sum_{i\in\boldsymbol{D}_{l,m}} t(\boldsymbol{x}_i) z_i y_i(1) + \frac{1}{n_m(1-a)}\sum_{i\in\boldsymbol{D}_{u,m}} t(\boldsymbol{x}_i) z_i y_i(1). \tag{22}$$

By assumptions (A1) and (A2),

$$\frac{1}{n_m}\sum_{i\in\boldsymbol{D}_{T,m}} w_{1i} z_i y_i(1) \overset{a.s.}{\to} E_{y,\boldsymbol{x},z}[y(1)zt(\boldsymbol{x})/e(\boldsymbol{x})] = E_{\boldsymbol{x}} E_{z|\boldsymbol{x}} E_{y|z,\boldsymbol{x}}[y(1)zt(\boldsymbol{x})/e(\boldsymbol{x})]$$
$$= E_x[E[y(1)|\boldsymbol{x}]t(\boldsymbol{x})] = \int E[(y(1)|\boldsymbol{x}]t(\boldsymbol{x})f(\boldsymbol{x})\Delta(d\boldsymbol{x}). \tag{23}$$

Since $|t(\boldsymbol{x}_i) z_i y_i(1)| \leq 1$, for all i, $\frac{1}{n_m a}\sum_{i\in\boldsymbol{D}_{l,m}} t(\boldsymbol{x}_i) z_i y_i(1) \overset{a.s.}{\to} 0$ if $n_{l,m}$ is finite, and $\frac{1}{n_m(1-a)}\sum_{i\in\boldsymbol{D}_{u,m}} t(\boldsymbol{x}_i) z_i y_i(1) \overset{a.s.}{\to} 0$ if $n_{u,m}$ is finite. When both $n_{l,m}$ and $n_{u,m}$ are infinite,

$$\frac{1}{n_m a}\sum_{i\in\boldsymbol{D}_{l,m}} t(\boldsymbol{x}_i) z_i y_i(1) = \frac{n_{l,m}}{n_m}\frac{1}{n_{l,m} a}\sum_{i\in\boldsymbol{D}_{l,m}} t(\boldsymbol{x}_i) z_i y_i(1) \overset{a.s.}{\to} 0, \tag{24}$$

as $\frac{1}{n_{l,m}}\sum_{i\in\boldsymbol{D}_{l,m}} t(\boldsymbol{x}_i) z_i y_i(1) \overset{a.s.}{\to} \int E[(y(1)|\boldsymbol{x}]e(\boldsymbol{x})t(\boldsymbol{x})f(\boldsymbol{x})\Delta(d\boldsymbol{x})$ and $\frac{n_{l,m}}{n_m} \to 0$ by (A2). Using the similar logic,

$$\frac{1}{n_m}\sum_{i\in\boldsymbol{D}_{u,m}} t(\boldsymbol{x}_i) z_i y_i(1) \overset{a.s.}{\to} 0. \tag{25}$$

From (22), (23), (24) and (25), we have

$$\frac{1}{n_m}\sum_{i=1}^{n_m} t(\boldsymbol{x}_i) z_i y_i(1) \to \int E[(y(1)|\boldsymbol{x}]t(\boldsymbol{x})f(\boldsymbol{x})\Delta(d\boldsymbol{x}). \tag{26}$$

Also,

$$\begin{aligned}\frac{1}{n_m}\sum_{i=1}^{n_m} w_{0i}(1-z_i)y_i(0) = &\frac{1}{n_m}\sum_{i\in \boldsymbol{D}_{T,m}} w_{0i}(1-z_i)y_i(0) + \frac{1}{n_m}\sum_{i\in \boldsymbol{D}_{l,m}} w_{0i}(1-z_i)y_i(0) \\ &+ \frac{1}{n_m}\sum_{i\in \boldsymbol{D}_{u,m}} w_{0i}(1-z_i)y_i(0).\end{aligned} \tag{27}$$

Following the similar arguments as above,

$$\begin{aligned}&\frac{1}{n_m}\sum_{i\in \boldsymbol{D}_{T,m}} w_{0i}(1-z_i)y_i(0) \overset{a.s.}{\to} E_{y,\boldsymbol{x},z}[y(0)(1-z)t(\boldsymbol{x})/(1-e(\boldsymbol{x}))] = \int E[(y(0)|\boldsymbol{x}]t(\boldsymbol{x})f(\boldsymbol{x})\Delta(d\boldsymbol{x}). \\ &\frac{1}{n_m}\sum_{i\in \boldsymbol{D}_{l,m}} w_{0i}(1-z_i)y_i(0) \overset{a.s.}{\to} 0, \quad \frac{1}{n_m}\sum_{i\in \boldsymbol{D}_{u,m}} w_{0i}(1-z_i)y_i(0) \overset{a.s.}{\to} 0\end{aligned} \tag{28}$$

We use the same argument as above to arrive at

$$\begin{aligned}&\frac{1}{n_m}\sum_{i=1}^{n_m} w_{1i} z_i \overset{a.s.}{\to} E_{y,\boldsymbol{x},z}[zt(\boldsymbol{x})/e(\boldsymbol{x})] = E_x[t(\boldsymbol{x})] = \int t(\boldsymbol{x})f(\boldsymbol{x})\Delta(d\boldsymbol{x}) \\ &\frac{1}{n_m}\sum_{i=1}^{n_m} w_{0i}(1-z_i) \overset{a.s.}{\to} E_{y,\boldsymbol{x},z}[(1-z)t(\boldsymbol{x})/(1-e(\boldsymbol{x}))] = E_x[t(\boldsymbol{x})] = \int t(\boldsymbol{x})f(\boldsymbol{x})\Delta(d\boldsymbol{x}).\end{aligned} \tag{29}$$

Hence, $\hat{\tau}_m^T \overset{a.s.}{\to} \left[\int\{E[y(1)|\boldsymbol{x}] - E[y(0)|\boldsymbol{x}]\}t(\boldsymbol{x})f(\boldsymbol{x})\Delta(d\boldsymbol{x})\right] / \left[\int t(\boldsymbol{x})f(\boldsymbol{x})\Delta(d\boldsymbol{x})\right] = \tau$ from (26), (28) and (29). Given that each $\hat{\tau}_m$ is bounded between -1 to 1, dominated convergence theorem leads to $E[\hat{\tau}_m^T] \to \tau$, as $n \to \infty$. Following the proof of Theorem 2 in Li *et al.* (2018), we have

$$\begin{aligned}n_m Var_{y|\boldsymbol{x},z}(\hat{\tau}_m^T) &= \frac{\frac{1}{n_m}\sum_{i=1}^{n_m} v_1(\boldsymbol{x}_i) z_i w_{1i}^2}{[\frac{1}{n_m}\sum_{i=1}^{n_m} z_i w_{1i}]^2} + \frac{\frac{1}{n_m}\sum_{i=1}^{n_m} v_0(\boldsymbol{x}_i)(1-z_i) w_{0i}^2}{[\frac{1}{n_m}\sum_{i=1}^{n_m}(1-z_i) w_{0i}]^2} \\ &= \frac{\frac{1}{n_m}\sum_{i=1}^{n_m} v_1(\boldsymbol{x}_i) z_i t(\boldsymbol{x}_i)^2/e^T(\boldsymbol{x}_i)^2}{[\frac{1}{n_m}\sum_{i=1}^{n_m} z_i t(\boldsymbol{x}_i)/e^T(\boldsymbol{x}_i)]^2} + \frac{\frac{1}{n_m}\sum_{i=1}^{n_m} v_0(\boldsymbol{x}_i)(1-z_i) t(\boldsymbol{x}_i)^2/(1-e^T(\boldsymbol{x}_i))^2}{[\frac{1}{n_m}\sum_{i=1}^{n_m}(1-z_i) t(\boldsymbol{x}_i)/(1-e^T(\boldsymbol{x}_i))]^2}.\end{aligned}$$

Note that

$$\frac{1}{n_m}\sum_{i=1}^{n_m} v_1(\boldsymbol{x}_i) z_i \frac{t(\boldsymbol{x}_i)^2}{e^T(\boldsymbol{x}_i)^2} = \frac{1}{n_m}\sum_{i\in \boldsymbol{D}_{T,m}} v_1(\boldsymbol{x}_i) z_i \frac{t(\boldsymbol{x}_i)^2}{e(\boldsymbol{x}_i)^2} + \frac{1}{n_m a^2}\sum_{i\in \boldsymbol{D}_{l,m}} v_1(\boldsymbol{x}_i) z_i t(\boldsymbol{x}_i)^2 + \frac{1}{n_m(1-a)^2}\sum_{i\in \boldsymbol{D}_{u,m}} v_1(\boldsymbol{x}_i) z_i t(\boldsymbol{x}_i)^2.$$

By assumptions (A1) and (A2), and using the arguments used before, $\frac{1}{n_m}\sum_{i\in \boldsymbol{D}_{l,m}} v_1(\boldsymbol{x}_i) z_i t(\boldsymbol{x}_i)^2 \overset{a.s.}{\to} 0$ and $\frac{1}{n_m}\sum_{i\in \boldsymbol{D}_{u,m}} v_1(\boldsymbol{x}_i) z_i t(\boldsymbol{x}_i)^2 \overset{a.s.}{\to} 0$, and $E_{\boldsymbol{x},z}[\frac{1}{n_m}\sum_{i\in \boldsymbol{D}_{T,m}} v_1(\boldsymbol{x}_i) z_i \frac{t(\boldsymbol{x}_i)^2}{e(\boldsymbol{x}_i)^2}] \to \int t(\boldsymbol{x})^2 \frac{v_1(\boldsymbol{x})}{e(\boldsymbol{x})} f(\boldsymbol{x})\Delta(\boldsymbol{x})$. Similarly,

$$\frac{1}{n_m}\sum_{i=1}^{n_m} v_0(\boldsymbol{x}_i) \frac{(1-z_i)t(\boldsymbol{x}_i)^2}{(1-e^T(\boldsymbol{x}_i))^2} = \frac{1}{n_m}\sum_{i\in \boldsymbol{D}_{T,m}} v_1(\boldsymbol{x}_i) \frac{(1-z_i)t(\boldsymbol{x}_i)^2}{(1-e^T(\boldsymbol{x}_i))^2} + \frac{1}{n_m(1-a)^2}\sum_{i\in \boldsymbol{D}_{l,m}} v_1(\boldsymbol{x}_i)(1-z_i) t(\boldsymbol{x}_i)^2 + \frac{1}{n_m a^2}\sum_{i\in \boldsymbol{D}_{u,m}} v_1(\boldsymbol{x}_i)(1-z_i) t(\boldsymbol{x}_i)^2.$$

Hence, $\frac{1}{n_m}\sum_{i\in \boldsymbol{D}_{l,m}} v_0(\boldsymbol{x}_i)(1-z_i) t(\boldsymbol{x}_i)^2 \overset{a.s.}{\to} 0$ and $\frac{1}{n_m}\sum_{i\in \boldsymbol{D}_{u,m}} v_0(\boldsymbol{x}_i)(1-z_i) t(\boldsymbol{x}_i)^2 \overset{a.s.}{\to} 0$, and $E_{\boldsymbol{x},z}[\frac{1}{n_m}\sum_{i\in \boldsymbol{D}_{T,m}} v_0(\boldsymbol{x}_i)(1-z_i) \frac{t(\boldsymbol{x}_i)^2}{(1-e(\boldsymbol{x}_i))^2}] \to \int t(\boldsymbol{x})^2 \frac{v_0(\boldsymbol{x})}{1-e(\boldsymbol{x})} f(\boldsymbol{x})\Delta(\boldsymbol{x})$. Using similar arguments,

$$\frac{1}{n_m}\sum_{i=1}^{n_m} z_i t(\boldsymbol{x}_i)/e(\boldsymbol{x}_i) \to \int t(\boldsymbol{x}) f(\boldsymbol{x})\Delta(\boldsymbol{x}), \;\; \frac{1}{n_m}\sum_{i=1}^{n_m}(1-z_i) t(\boldsymbol{x}_i)/(1-e(\boldsymbol{x}_i)) \to \int t(\boldsymbol{x}) f(\boldsymbol{x})\Delta(\boldsymbol{x}).$$

Using Slutsky's theorem,

$$n_m E_{\boldsymbol{x}}[Var(\hat{\tau}_m^T|\boldsymbol{x})] \to \int t(\boldsymbol{x})^2 \left\{\frac{v_1(\boldsymbol{x})}{e(\boldsymbol{x})} + \frac{v_0(\boldsymbol{x})}{(1-e(\boldsymbol{x}))}\right\} f(\boldsymbol{x})\Delta(\boldsymbol{x}) / \left\{\int t(\boldsymbol{x}) f(\boldsymbol{x})\Delta(\boldsymbol{x})\right\}^2.$$

As $n_m \to \infty$ $E_{\boldsymbol{x}}[Var(\hat{\tau}_m^T|\boldsymbol{x})] \to 0$. Following Imbens (2004), typically, $Var_{\boldsymbol{x}}(E[\hat{\tau}_m^T|\boldsymbol{x})]) \leq E_{\boldsymbol{x}}[Var(\hat{\tau}_m^T|\boldsymbol{x})] \to 0$. Hence, $Var_{y,\boldsymbol{x},z}(\hat{\tau}_m^T) \to 0$.

Combining equations (19), (20), (21) and the result above, as $n \to \infty$

$$P(|\bar{\tau} - \tau| > c) \leq 2\exp(-M\epsilon(1-\pi)c/6). \tag{30}$$

which concludes the proof of (i).

To prove (ii), note that

$$P(|\bar{V} - V| > c) \leq P(|\bar{V} - \bar{V}^{T,\epsilon}| > c/3) \\ + P(|\bar{V}^T - \bar{V}^{T,\epsilon}| > c/3) + P(|\bar{V}^T - V| > c/3).$$

The first and second terms are straightforward to bound following the Laplace distribution. More specifically,

$$P(|\bar{V}^T - \bar{V}^{T,\epsilon}| > c/3) = E_{y,\boldsymbol{x},z} P(|\bar{V}^T - \bar{V}^{T,\epsilon}| > c/3|\boldsymbol{D}) = \exp(-Mc\epsilon\pi/6)$$
$$P(|\bar{V} - \bar{V}^{T,\epsilon}| > c/3) = E_{y,\boldsymbol{x},z} P(|\bar{V} - \bar{V}^{T,\epsilon}| > c/3|\boldsymbol{D}) \leq \exp(-Mc\epsilon\pi/6). \quad (31)$$

Regarding the third term, note that $\bar{V}^T = \sum_{m=1}^{M} \hat{V}_m^T/M$ and $\hat{V}_m^T \xrightarrow{P} V$ as $n_m \to \infty$, using the above results. Hence, $\bar{V}^T \xrightarrow{P} V$ which implies $P(|\bar{V}^T - V| > c/3) \to 0$ as $n \to \infty$. Hence, $P(|\bar{V}^T - V| > c/3) \leq 2\exp(-M\epsilon\pi c/6)$ as $n \to \infty$, proving (ii).

7.2 Proof of Theorem 3.5

We have

$$KL(\tilde{g}^\epsilon, g) = \frac{(\bar{\tau} - \tau)^2}{2V} + \frac{\bar{V}}{2V} - \frac{1}{2} - \frac{1}{2}\log\frac{\bar{V}}{V} = U_1 + U_2 + U_3,$$

where $U_1 = \frac{(\bar{\tau}-\tau)^2}{2V}$, $U_2 = \frac{\bar{V}}{2V} - \frac{1}{2}$ and $U_3 = -\frac{1}{2}\log\frac{\bar{V}}{V}$. Following Lemma 3.4, for any $c > 0$, as $n \to \infty$,

$$P(|U_1| > c/3) = P\left(\frac{(\bar{\tau} - \tau)^2}{2V} > c/3\right) \leq 2\exp\left(-\frac{M\epsilon(1-\pi)\sqrt{2Vc}}{6\sqrt{3}}\right) \quad (32)$$
$$P(|U_2| > c/3) = P\left(\left|\frac{\bar{V}}{2V} - \frac{1}{2}\right| > c/3\right) \leq 2\exp(-M\epsilon\pi Vc/9) \quad (33)$$
$$P(|U_3| > c/3) = P\left(\left|\frac{1}{2}\log\frac{\bar{V}}{V}\right| > c/3\right) \\ \leq P\left(\left|\frac{\bar{V}}{2V} - \frac{1}{2}\right| > c/3\right) \leq 2\exp(-M\epsilon\pi Vc/9), \quad (34)$$

where last inequality uses the fact that $log(h) \leq h - 1$, for any $h > 0$. Finally, for any $c > 0$,

$$\begin{aligned} P(KL(\tilde{g}^{\epsilon}, g) > c) &\leq P(|U_1| > c/3) + P(|U_2| > c/3) + P(|U_3| > c/3) \\ &\leq 2\exp\left(-\frac{M\epsilon(1-\pi)\sqrt{2Vc}}{6\sqrt{3}}\right) + 4\exp(-M\epsilon\pi Vc/9), \end{aligned}$$

as $n \to \infty$, proving the result.

References

Abadi, M., Chu, A., Goodfellow, I., McMahan, H. B., Mironov, I., Talwar, K., and Zhang, L. (2016). Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, pages 308–318.

Balle, B., Barthe, G., Gaboardi, M., Hsu, J., and Sato, T. (2020). Hypothesis testing interpretations and renyi differential privacy. In *International Conference on Artificial Intelligence and Statistics*, pages 2496–2506. PMLR.

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

Becker, B. and Kohavi, R. (1996). Adult. UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C5XW20.

Blanco-Justicia, A., Sanchez, D., Domingo-Ferrer, J., and Muralidhar, K. (2022). A critical review on the use (and misuse) of differential privacy in machine learning. *ACM Computing Surveys*, **55**(8), 1–16.

Crump, R. K., Hotz, V. J., Imbens, G. W., and Mitnik, O. A. (2009). Dealing with limited overlap in estimation of average treatment effects. *Biometrika*, **96**(1), 187–199.

D'Orazio, V., Honaker, J., and King, G. (2015). Differential privacy for social science inference. Sloan Foundation Economics Research Paper No. 2676160.

Dwork, C. (2006). Differential privacy. In *Automata, Languages and Programming: 33rd International Colloquium, ICALP 2006, Venice, Italy, July 10-14, 2006, Proceedings, Part II 33*, pages 1–12. Springer.

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

Dwork, C., Smith, A., Steinke, T., and Ullman, J. (2017). Exposed! a survey of attacks on private data. *Annual Review of Statistics and Its Application*.

Fang, X., Yu, F., Yang, G., and Qu, Y. (2019). Regression analysis with differential privacy preserving. *IEEE Access*, **7**, 129353–129361.

Gaboardi, M., Rogers, R., and Sheffet, O. (2019). Locally private mean estimation: z-test and tight confidence intervals. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 2545–2554. PMLR.

Grzybowski, M., Clements, E. A., Parsons, L., Welch, R., Tintinalli, A. T., Ross, M. A., and Zalenski, R. J. (2003). Mortality benefit of immediate revascularization of acute st-segment elevation myocardial infarction in patients with contraindications to thrombolytic therapy: a propensity analysis. *JAMA*, **290**(14), 1891–1898.

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

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

Ji, Z., Lipton, Z. C., and Elkan, C. (2014). Differential privacy and machine learning: a survey and review. *arXiv preprint arXiv:1412.7584*.

Kang, J. D. and Schafer, J. L. (2007). Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. *Statistical Science*, **22**(4), 523–539.

Kurth, T., Walker, A. M., Glynn, R. J., Chan, K. A., Gaziano, J. M., Berger, K., and Robins, J. M. (2006). Results of multivariable logistic regression, propensity matching, propensity adjustment, and propensity-based weighting under conditions of nonuniform effect. *American Journal of Epidemiology*, **163**(3), 262–270.

Lee, S. K., Gresele, L., Park, M., and Muandet, K. (2019). Privacy-preserving causal inference via inverse probability weighting. arXiv:1905.12592.

Li, F., Morgan, K. L., and Zaslavsky, A. M. (2018). Balancing covariates via propensity score weighting. *Journal of the American Statistical Association*, **113**(521), 390–400.

Mivule, K., Turner, C., and Ji, S.-Y. (2012). Towards a differential privacy and utility preserving machine learning classifier. *Procedia Computer Science*, **12**, 176–181.

Nishihara, R., Murray, I., and Adams, R. P. (2014). Parallel mcmc with generalized elliptical slice sampling. *Journal of Machine Learning Research*, **15**(1), 2087–2112.

Nissim, K., Raskhodnikova, S., and Smith, A. (2007). Smooth sensitivity and sampling in private data analysis. In *Proceedings of the Thirty-ninth Annual ACM Symposium on Theory of Computing*, pages 75–84.

Niu, F., Nori, H., Quistorff, B., Caruana, R., Ngwe, D., and Kannan, A. (2022). Differentially private estimation of heterogeneous causal effects. *Proceedings of Machine Learning Research*, **140**, 1—-17.

Ohnishi, Y. and Awan, J. (2023). Locally private causal inference. arxiv:2301.01616.

Pensia, A., Asadi, A. R., Jog, V., and Loh, P.-L. (2023). Simple binary hypothesis testing under local differential privacy and communication constraints. In *The Thirty Sixth Annual Conference on Learning Theory*, pages 3229–3230. PMLR.

Rosenbaum, P. R. and Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, **70**(1), 41–55.

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

Rubin, D. B. (1987). *Multiple Imputation for Survey Nonresponse*. New York: Wiley.

Triastcyn, A. and Faltings, B. (2020). Bayesian differential privacy for machine learning. In *International Conference on Machine Learning*, pages 9583–9592. PMLR.

Vincent, J. L., Baron, J.-F., Reinhart, K., Gattinoni, L., Thijs, L., Webb, A., Meier-Hellmann, A., Nollet, G., Peres-Bota, D., investigators, A., *et al.* (2002). Anemia and blood transfusion in critically ill patients. *JAMA*, **288**(12), 1499–1507.

Wang, Y., Si, C., and Wu, X. (2015). Regression model fitting under differential privacy and model inversion attack. In *Proceedings of the 24th International Conference on Artificial Intelligence*, pages 1003–1009.

Zhang, J., Zhang, Z., Xiao, X., Yang, Y., and Winslett, M. (2012). Functional mechanism: Regression analysis under differential privacy. *arXiv preprint arXiv:1208.0219*.

Zheng, H., Hu, H., and Han, Z. (2020). Preserving user privacy for machine learning: Local differential privacy or federated machine learning? *IEEE Intelligent Systems*, **35**(4), 5–14.

Zhou, X. and Reiter, J. P. (2010). A note on Bayesian inference after multiple imputation. *The American Statistician*, **64**, 159–163.

	η	2	2	2	4	4	4
	γ	0	1	2	0	1	2
ATE	Avg. τ_{ATE}	0	.204	.342	0	.204	.343
	$\hat{\tau}_{ATE}$						
	RMSE	.011	.014	.012	.019	.019	.022
	95% CI coverage	90.2	89.8	90.8	91.2	92.4	92.0
	95% CI length	.054	.058	.057	.171	.192	.187
	$\tilde{\tau}^{\epsilon}_{ATE}$						
	RMSE	.016	.016	.015	.023	.021	.024
	95% CI coverage	96.8	97.4	97.2	98.0	98.0	98.2
	95% CI length	.113	.134	.148	.281	.295	.316
ATT	Avg. τ_{ATT}	0	.205	.345	0	.206	.348
	$\hat{\tau}_{ATT}$						
	RMSE	.012	.012	.010	.022	.017	.019
	95% CI coverage	89.8	90.8	91.6	91.0	91.8	92.0
	95% CI length	.059	.062	.062	.201	.229	.213
	$\tilde{\tau}^{\epsilon}_{ATT}$						
	RMSE	.016	.014	.013	.026	.023	.021
	95% CI coverage	96.6	96.8	97.4	97.0	98.2	98.0
	95% CI length	.136	.144	.169	.303	.324	.332
ATC	Avg. τ_{ATC}	0	.202	.338	0	.202	.337
	$\hat{\tau}_{ATC}$						
	RMSE	.015	.014	.010	.022	.020	.017
	95% CI coverage	90.2	90.4	91.0	91.6	92.2	92.4
	95% CI length	.057	.065	.064	.215	.234	.219
	$\tilde{\tau}^{\epsilon}_{ATC}$						
	RMSE	.018	.018	.015	.028	.025	.026
	95% CI coverage	97.2	97.4	97.4	98.0	98.2	98.4
	95% CI length	.132	.159	.179	.310	.326	.341

Table 2: Results from 500 simulations with ($M = 100, \epsilon = 1, a = .05$) for the ATE, ATT and ATC. Results include the root of the average squared errors (RMSE) between the differentially private WATE estimate and the corresponding true value based on (16) – (18); the percentage of the five hundred 95% confidence intervals that cover the corresponding treatment effect; and, the average length of the estimated 95% confidence intervals in parenthesis. These quantities are shown for both the privacy protected and non-privacy protected estimation.

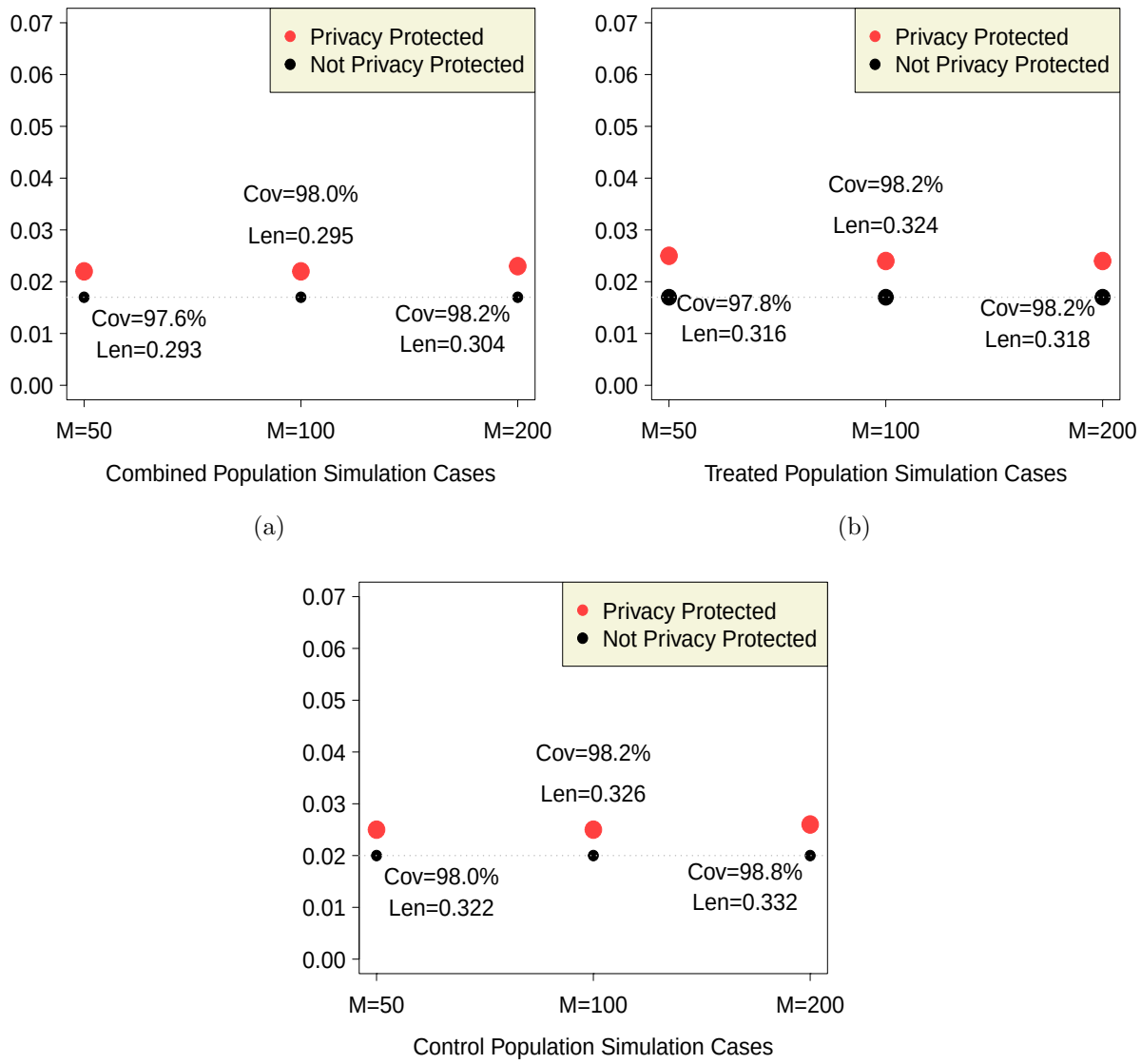


Figure 2: Root mean squared error (RMSE) of the differentially private WATEs for the ATE (Figure 2(a)), ATT (Figure 2(b)), and ATC (Figure 2(c)), for $M \in \{50, 100, 200\}$. In all cases, $n = 10000$, $\epsilon = 1$ and $a = .05$. RMSEs for WATE estimates with no privacy protection are denoted by black dots. RMSEs for the differentially private WATE estimates are denoted by red dots. "Cov" and "Len" stand for the coverage rate and average length of the 95% intervals, both based on the differentially private algorithms.

	$\text{Dev}(\tilde{\tau}^{\epsilon,a}_{ATE})$	$\text{Dev}(\tilde{\tau}^{\epsilon,a}_{ATT})$	$\text{Dev}(\tilde{\tau}^{\epsilon,a}_{ATC})$	$\text{Cov}(\tilde{\tau}^{\epsilon,a}_{ATE})$	$\text{Cov}(\tilde{\tau}^{\epsilon,a}_{ATT})$	$\text{Cov}(\tilde{\tau}^{\epsilon,a}_{ATC})$
$a = 0.03$	0.005	0.006	0.005	0.984	0.986	0.988
$a = 0.07$	0.005	0.006	0.006	0.978	0.978	0.982
$a = 0.10$	0.007	0.008	0.008	0.974	0.976	0.978

Table 3: Results for simulations for $a \in \{0.03, 0.07, 0.10\}$. Entries include the average absolute difference in the differentially private WATE and the non-private WATE without truncation, labeled $\text{Dev}(\tilde{\tau}^{\epsilon,a})$, and the coverage rate of the 95% intervals, labeled $\text{Cov}(\tilde{\tau}^{\epsilon,a})$. In all cases, $M = 100$, $n = 10000$, and $\epsilon = 1$.

	$n = 100000$			$n = 5000$		
	Combined	Treated	Control	Combined	Treated	Control
$\hat{\tau}$						
RMSE	0.012	0.013	0.013	0.022	0.023	0.025
95% CI coverage	0.928	0.918	0.920	0.970	0.972	0.972
95% CI length	0.165	0.189	0.184	0.291	0.345	0.343
$\tilde{\tau}^{\epsilon}$						
RMSE	0.017	0.018	0.018	0.028	0.029	0.031
95% CI coverage	0.972	0.976	0.978	0.996	0.998	0.998
95% CI length	0.254	0.271	0.276	0.648	0.729	0.742

Table 4: Results with sample sizes $n = 100000$ and $n = 5000$, with $(M = 100, \epsilon = 1, a = .05)$. Here, $\tilde{\tau}^{\epsilon}$ is the privacy protected treatment effect estimate. Results include coverage rates and average lengths of the 95% intervals.

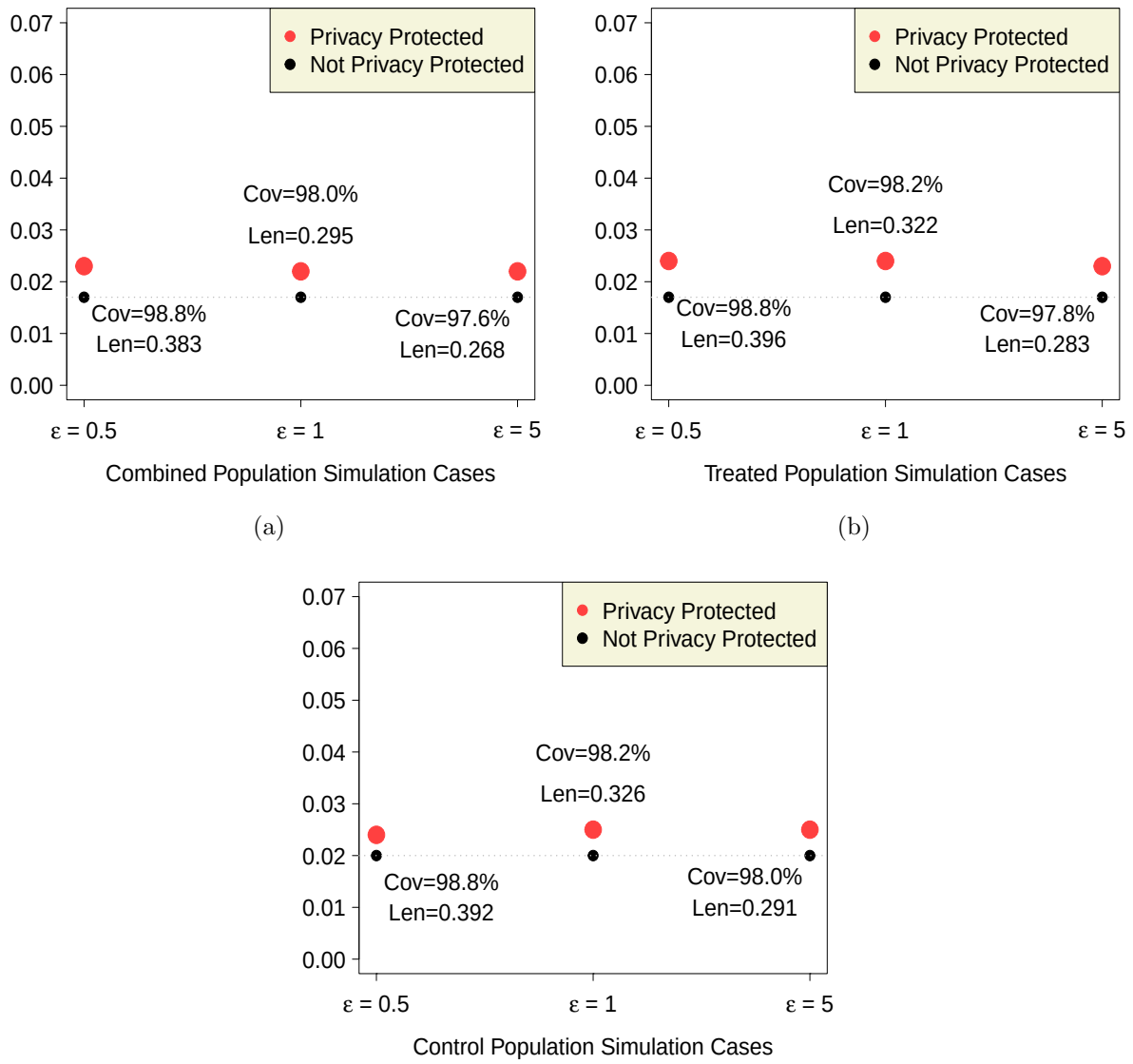


Figure 3: Root mean squared error (RMSE) of the differentially private WATEs for the ATE (Figure 3(a)), ATT (Figure 3(b)), and ATC (Figure 3(c)) for $\epsilon \in \{0.5, 1, 5\}$. In all these cases, $M = 100$, $n = 10000$, and $a = 0.05$. RMSEs for the WATE estimates with no privacy protection are denoted by black dots. RMSEs for the differentially private WATE estimates are denoted by red dots. "Cov" and "Len" stand for the coverage rate and average length of the 95% intervals, both based on the differentially private algorithms.

	ϵ	ATE 95% CI	ATT 95% CI	ATC 95% CI
$\tilde{\tau}^{\epsilon}$	1	0.271	0.258	0.236
		(0.183, 0.361)	(0.170, 0.344)	(0.149, 0.324)
	0.5	0.272	0.269	0.275
		(0.144, 0.401)	(0.136, 0.406)	(0.151, 0.407)
$\hat{\tau}$		0.263	0.271	0.260
		(0.251, 0.275)	(0.259, 0.283)	(0.248, 0.272)

Table 5: Point estimates and 95% intervals for treatment effects from the analysis of the Adult Income Data described in Section 5. Results include the differentially private inferences based on Algorithm 1 in the panel indicated by $\tilde{\tau}^{\epsilon}$, as well as results based on the full data with no privacy protection or truncation in the panel headed by $\hat{\tau}$. Differentially private results use $M = 100$ and $a = 0.05$.