Causal Inference under Differential Privacy: Challenges and Mitigation Strategies

Amirhossein Farzam
Duke University
a.farzam@duke.edu

Guillermo Sapiro
Princeton University & Duke University & Apple
guillermos@princeton.edu

Abstract

The intersection of differential privacy (DP) and causal inference is crucial for protecting data privacy while preserving the accuracy of causal estimates, yet limited research explores how DP mechanisms impact causal effects estimated from privatized data. This paper formally investigates for the first time the impact of DP on causal inference, focusing on how standard DP mechanisms affect the accuracy of treatment effect estimates across various causal inference frameworks. Our core theoretical findings reveal that while applying DP mechanisms to outcomes preserves the unbiasedness of average treatment effect (ATE) estimates, it can increase their variance and severely distort individual treatment effect (ITE) estimates. Privatizing treatments, on the other hand, can lead to unexpected consequences, such as ITE underestimation in certain settings. Informed by our theoretical analysis, we propose two solutions to these issues: First, we show that under some conditions with privatized treatments, we can exactly recover the ATE estimates given information about the true privacy and balance parameters. Second, we propose robust regression as a mitigation strategy during the data analysis stage, to reduce the ITE estimation error. In addition to providing actionable guidance for balancing data privacy and causal inference accuracy, this work provides the first foundational results on treatment effect estimation from differentially private data, laying the foundation for future research into privacy-aware causal representation learning toward enabling models to robustly handle privatized data.

1 Introduction

Differential privacy (DP) has emerged as the leading approach for safeguarding individual privacy across areas such as healthcare, finance, social sciences, and machine learning (Dwork et al., 2006; Dwork & Roth, 2014). By limiting the amount of information revealed about any single data point, DP provides a mathematically sound framework for reducing privacy risks in sensitive datasets. Meanwhile, causal inference is essential for understanding the effects of interventions, guiding policy decisions, and advancing research in fields like medicine and social sciences, where data privacy is a critical concern (Pearl, 2009; Imbens & Rubin, 2015). The integration of differential privacy into causal inference presents both significant opportunities and challenges, with important implications for causal representation learning. Ensuring that causal inference on non-private data produces private estimates is an important and challenging problem (Kusner et al., 2016; Ohnishi & Awan, 2023; Niu et al., 2022; Guha & Reiter, 2024; Rho et al., 2023), and conversely, causal estimates derived from privatized data can be distorted by the noise introduced to meet DP requirements (Javanmard et al., 2023; Agarwal & Singh, 2021). As the use of sensitive data for causal analysis grows, it is crucial to understand how privacy-preserving mechanisms impact causal estimates as well as causal representation learning (Schölkopf, 2022; Schölkopf et al., 2021). In this paper, we systematically examine the effects of differential privacy on causal inference. We not only identify the problem but also analyze the challenges posed by standard DP mechanisms in estimating causal effects,

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

offering guidance for downstream analysis and proposing mitigation strategies. Our work addresses the theoretical implications of DP and offers practical guidelines for applying it in standard causal inference settings, laying the groundwork for future research on causal representation learning from private data.

Despite the growing body of work on DP and causal inference, a significant gap remains in understanding how standard DP mechanisms affect the accuracy of causal estimates. While prior research has focused on developing privacy mechanisms for causal inference (Kusner et al., 2016; Ohnishi & Awan, 2023; Niu et al., 2022; Guha & Reiter, 2024; Rho et al., 2023), little attention has been paid to how privatized data impacts downstream causal inference tasks (Javanmard et al., 2023; Agarwal & Singh, 2021). Addressing this gap, we conduct a theoretical analysis of the error induced by DP in treatment effect estimates across various causal inference frameworks, including Randomized Controlled Trials (RCTs) (Fisher, 1935), Difference-in-Differences (DiD) (Angrist & Pischke, 2009; Ashenfelter & Card, 1984), and matching (Gu & Rosenbaum, 1993; Imbens, 2004; Abadie & Imbens, 2016). To our knowledge, this paper provides the first comprehensive analysis of how DP impacts causal inference across a range of standard conditions, setting the stage for future research on causal representation learning in privacy-sensitive environments. We examine how treatment effect estimates derived from differentially private data deviate from those obtained using true (non-private) data, focusing on the impact of DP on key causal quantities such as the Average Treatment Effect (ATE) and Individual Treatment Effect (ITE). Our results allow us to provide actionable guidance for downstream data analysis, which paves the path for developing robust methods for causal representation learning from private data.

Main contributions This paper provides a comprehensive analysis of how differential privacy impacts key causal estimates, like the ATE and ITE, in standard causal inference frameworks, highlighting conditions for their reliability. We propose strategies to recover the exact treatment effect estimates in some settings, and to reduce DP-induced error during data analysis in others. Along with offering actionable guidance, our findings lay the groundwork for future research on privacy-aware causal representation learning that can handle privatized data.

2 Preliminaries

Differential privacy Differential privacy (DP) is a framework that ensures the inclusion or exclusion of a single data point does not significantly impact the result of any analysis, providing strong privacy guarantees (Dwork et al., 2006). A randomized algorithm $\mathcal{A}$ is (ϵ, δ)-differentially private if, for all datasets D and D' differing in one element,

$$\mathbb{P}[\mathcal{A}(D) \in S] \leq e^{\epsilon}\mathbb{P}[\mathcal{A}(D') \in S] + \delta. \tag{1}$$

where S is a subset of the output space (Dwork & Roth, 2014). Common DP mechanisms include the Laplace and Gaussian mechanisms for continuous, and randomized response for binary data. To achieve ϵ-differential privacy, noise proportional to the function's sensitivity, $\Delta f = \max_{D,D'} \|f(D) - f(D')\|_1$, is added to the output. The Laplace mechanism adds a Laplace noise with scale parameter $\lambda = \frac{\Delta f}{\epsilon}$, while the Gaussian mechanism adds noise based on a variance proportional to the privacy budget ϵ and sensitivity Δf. Randomized response flips binary values with a certain probability. These mechanisms lead to a trade off between privacy and accuracy.

Causal inference Causal inference seeks to estimate the effect of a treatment T on an outcome Y, given covariates X, which can be quantified at both the individual and population levels as the individual treatment effect (ITE) and the average treatment effect (ATE), respectively (Rubin, 2005). The ITE for individual i and the ATE for the population are defined as:

$$\tau_i = \mathbb{E}\left[Y_i | do(T = t_i) - Y_i | do(T = {t_i}')\right] \qquad \tau = \mathbb{E}\left[Y | do(T = t) - Y | do(T = t')\right], \tag{2}$$

where $do(\cdot)$ represents an intervention imposing treatment t (Pearl, 2009). The challenge in estimating these effects arises from the unobservability of counterfactual outcomes. To address this, key assumptions are required, such as those made for randomized controlled trials (RCTs) (Fisher, 1935) and difference-in-differences (DiD) (Angrist & Pischke, 2009). RCTs provide unbiased ATE estimates through randomized treatment assignment, while DiD compares outcome changes over time between treated and control groups in observational studies.

3 Related work

Related work on differential privacy and causal inference is discussed throughout the paper, with a more detailed review in Appendix A. While many studies have investigated mechanisms for private causal inference, to our knowledge, a systematic analysis of how DP mechanisms impact causal inference from privatized data, as presented here, has not been explored.

4 Causal inference with private outcome

We examine how applying differential privacy to the outcome variable Y affects ATE and ITE estimates in RCT and DiD settings by analyzing the impact of Laplace and Gaussian mechanisms on the accuracy of these estimates. Note that our analysis in RCT settings directly extends to regression discontinuity design with uniform kernel around the discontinuity point as well, making it relevant to some of the most popular tools in causal inference, in addition to providing the framework and tools for additional ones. In each case, we aim to assess the effect of privacy-induced noise on causal inference and offer strategies to mitigate these errors. Let $\tilde{Y}$ denote the privatized outcome. We denote the conditional expectations of Y by $\mu_1 = \mathbb{E}[Y \mid T = 1]$ and $\mu_0 = \mathbb{E}[Y \mid T = 0]$, and their private counterparts by $\tilde{\mu}_1$ and $\tilde{\mu}_0$. We calculate the ATE and ITE from $\tilde{Y}$, denoted $\tilde{\tau}$ and $\tilde{\tau}_i$, and define the corresponding errors as:

$$Z := \tilde{\tau} - \tau, \qquad\qquad Z_i := \tilde{\tau}_i - \tau_i, \tag{3}$$

where τ and τ_i are the non-private ATE and ITE. For simplicity, we analyze $Z > 0$, noting that the extension to $Z < 0$ is straightforward. We use $\bar{\cdot}$ to denote sample equivalents of population values.

4.1 ATE with private outcome

When DP is applied to the outcome variable Y in an RCT, the ATE is estimated as the difference in expected outcomes between the treatment and control groups, $\tau = \mu_1 - \mu_0$. With an ϵ-differentially private outcome using the Laplace mechanism yielding $\tilde{Y} = Y + \text{Lap}(\lambda)$, where $\lambda = \frac{\Delta f}{\epsilon}$, the ATE estimate remains unbiased, while its variance increases as stated in the following proposition.

Proposition 4.1. *Given a sample with n_1 individuals in the treatment and n_0 in the control groups, we have $\tau = \tilde{\tau}$ and $\mathit{Var}(\bar{\tilde{\tau}}) = \mathit{Var}(\bar{\tau}) + 2\lambda^2 \left(\frac{1}{n_1} + \frac{1}{n_0}\right)$.*

The proof and further details are provided in appendices B and C. For Gaussian noise, the additional variance term is $\sigma^2 \left(\frac{1}{n_1} + \frac{1}{n_0}\right)$. Results are analogous in the DiD setting, since the estimator is essentially equivalent to the difference between two estimators in the RCT case, adding two additional sample size terms to the variance, while the mean remains unaffected. For details, see Appendix B.

Implications and mitigation. The statement of Propositions 4.1 indicates that average causal effects remain unbiased upon applying common DP mechanisms to the outcomes. However, DP leads to an increase in sample variance, which can be reduced by collecting a larger sample. The derived formulas provide explicit indication on how larger the sample is needed to mitigate the variance increase.

4.2 ITE with private outcome

In the absence of averaging, the noise in $\tilde{Y}$ does not cancel out, leading to non-zero errors in ITE estimates. Due to the fundamental problem of causal inference, ITE cannot be directly computed from data. However, under standard assumptions, a variety of methods could be used to estimate ITE via inferring potential outcomes or matching (Imbens & Rubin, 2015; Shalit et al., 2017). Hence, we consider the potential outcomes $\tilde{Y}_i \mid do(T = 1)$ and $\tilde{Y}_i \mid do(T = 0)$ for this analysis. See Appendix B for further details. For RCTs, with $\tilde{Y} = Y + \text{Lap}(\lambda)$, the ITE is given by $\tilde{\tau}_i = \tau_i + \xi_1 - \xi_0$, where ξ_1 and ξ_0 are independent Laplace random variables. The resulting error distribution is stated in Proposition 4.2.

Proposition 4.2. *The probability density function of the ITE error Z_i is given by $f_{Z_i}(z) = \frac{1}{2\lambda} e^{-\frac{z}{\lambda}} \left(1 + \frac{z}{\lambda}\right)$, yielding an expected error of $\frac{3}{2}\lambda^2$.*

Here for simplicity we assume that both treatment and control groups are privatized using the same noise distribution parameter. In Aappendix C, we extend this analysis to the more general case of different noise parameters for the two groups, which is particularly relevant in situations where the sensitivity of the outcome function is dependent on the treatment assignment. With the Gaussian mechanism, the error is the subtraction of two Gaussian noises, and hence $Z_i \sim \mathcal{N}(0, 2\sigma^2)$. In DiD settings, the ITE can be written as $\tilde{\tau}_i = \tau_i + (\xi_{1,t} - \xi_{1,t-1}) + (\xi_{0,t} - \xi_{0,t-1})$, essentially the difference between two ITE errors in the RCT case. This leads to the following critical proposition.

Proposition 4.3. *The expected error of ITE estimated from $\tilde{Y}$ with a DiD approach diverges.*

Similar to the RCT case, for the Gaussian mechanism, DiD involves four Gaussian noises, resulting in $Z_i \sim \mathcal{N}(0, 4\sigma^2)$. For more details and proofs, see appendices B and C.

Implications and mitigation. The error distribution in Proposition 4.2, allows us to assess the ITE error in RCT settings in details. As we discuss in Section 6, using a robust regression tends to reduce the average ITE error. Proposition 4.3 on the other hand, indicates that estimating individual-level causal effects using ITE in a DiD framework is problematic since its expected error is infinite.

5 Causal inference with private treatment

In this section, we explore the impact of differential privacy on the treatment variable T, using the randomized response mechanism for a binary treatment. This mechanism flips the treatment assignment with probability $1 - p$, where $p := \mathbb{P}[\tilde{T} = T]$ is the probability of reporting the true treatment, which we assume is better than random, i.e., $p > \frac{1}{2}$. We analyze its effects on ATE estimation in RCT settings and ITE in matching-based approaches. Let $q := \mathbb{P}[T = 1]$ denote the proportional size of the treatment group. Other notation follows Section 4.

5.1 ATE with private treatment

When the treatment assignment T is privatized using the randomized response mechanism, we can express $\tilde{\mu}_1$ and $\tilde{\mu}_0$ in terms of μ_1, μ_0, p, and q. In an RCT setting, this leads to the following result, as we show in Appendix D.

Proposition 5.1. *The expected private outcomes can be expressed in terms of μ_0 and μ_1:*

$$\tilde{\mu}_1 = \alpha_1 \mu_1 + \beta_1 \mu_0 \qquad\qquad \tilde{\mu}_0 = \alpha_0 \mu_1 + \beta_0 \mu_0, \tag{4}$$

where α_1, β_1, α_0, and β_0 are constant functions of p and q, as we detail in Appendix D.

This proposition gives a system of linear equations, which we can solve to exactly recover the true values of the conditional expected outcomes μ_1 and μ_0, and consequently the true ATE. Solving this system yields

$$\mu_1 = \frac{\tilde{\mu}_0 - \frac{\beta_0}{\beta_1}\tilde{\mu}_1}{\alpha_0 - \frac{\beta_0}{\beta_1}\alpha_1} \qquad\qquad \mu_0 = \frac{\tilde{\mu}_1 - \alpha_1 \mu_1}{\beta_1}, \tag{5}$$

which we can use to find the true ATE in an RCT setting as well as the exact value of the error $Z = (\tilde{\mu}_1 - \tilde{\mu}_0) - (\mu_1 - \mu_0)$.

Implications and mitigation. An important implication of Proposition 5.1 is due to Equation 4 forming a system of linear equations for μ_0 and μ_1, which we can solve to exactly recover the true ATE. Furthermore, using these solutions, we can express Z in terms of $\tilde{\mu}_1$, $\tilde{\mu}_0$, p, q, which allows us to explore how the error changes with respect to the privacy parameter p, balance parameter q, and the observed conditional outcomes.

5.2 ITE with private treatment

As noted in Section 4.2, several methods exist for inferring ITE from data, with one popular approach being matching (Gu & Rosenbaum, 1993; Imbens, 2004; Abadie & Imbens, 2016). We provide the analysis of the error in estimated ITE due to privatizing the treatments considering a matching approach; other methods can be analyzed following similar approaches. We denote the ITE estimated by matching individuals i and j from the treated and control groups by $\tau_{i\sim j}$, and we extend this notation to other variables; for instance, $\tilde{\tau}_{i\sim j}$ and $Z_{i\sim j}$. In this setup, with treatment values privatized via randomized response, we find the following result (derived in Appendix D).

Proposition 5.2. *The expected ITE estimated using $\tilde{T}$, and the expected ITE estimation error are*

$$\mathbb{E}\left[\tilde{\tau}_{i\sim j}\right] = p\left(\tau_i + \tau_j\right) - \tau_{j\sim i}, \tag{6}$$

$$\mathbb{E}\left[Z_{i\sim j}\right] = \left(\tau_i + \tau_j\right)\left(1-p\right). \tag{7}$$

Note that the results stated in Proposition 5.2 are based on the true ITE values τ_i and τ_j computed from potential outcomes, and $\tau_{j\sim i}$ from true data. Since we cannot compute any of these quantities from privatized data, we cannot manipulate the ITE estimation error in this case. However, this result helps us better understand the expected error, as discussed next.

Implications. The expression in Equation 7 indicates that a larger p leads to a smaller ITE estimation error. This is anticipated, as a larger p implies a smaller error in $\tilde{T}$. A less predictable observation, implied from the expression in Equation 6, is that the expected estimated ITE becomes larger as we increase p. These two observations together lead to an important consequence: the ITE estimates using privatized treatment are consistently smaller than the ITE estimated from the true data.

6 Robust regression for reducing the treatment effect estimation error

In Section 5.1, we showed that when the treatments are privatized using the randomized response mechanism, we can exactly recover the non-private ATE estimates in an RCT setting, given information about the privacy parameter and the balance in the size of the treatment and control groups. In this section, we propose a mitigation strategy to reduce the DP-induced error in causal effect estimates which is applicable and effective at the data analysis stage, without any information about the raw data. Informed by the analysis in Section 4, we show that robust regression can reduce ITE estimation error when outcomes are privatized with the Laplace mechanism. Further details are provided in Appendix E.

When the outcome variable is privatized using the Laplace mechanism, the fat-tailed nature of the noise introduces outliers, affecting ITE estimation accuracy, especially with standard linear regression (Rousseeuw & Leroy, 2005). To mitigate this, we propose the use of robust regression, which is designed to mitigate the impact of outliers on the regression model. We use the Theil-Sen estimator (Theil, 1950; Sen, 1968; Akritas et al., 1995), which estimates the slope of the regression line based on the median of the slopes of lines passing through pairs of points in the dataset. The Theil-Sen estimator is less sensitive to outliers than ordinary least squares (OLS) regression, while, unlike more aggressive robust regression methods such as *random sample consensus* (RANSAC) (Fischler & Bolles, 1981), it still incorporates extreme values into the analysis. This makes it well-suited for privatized data, where both genuine extremes and noise-induced outliers exist. In RCT settings, applying Theil-Sen regression to privatized outcome data significantly reduces ITE estimation error when outcome values are inferred using a linear regression model (Figure 2).

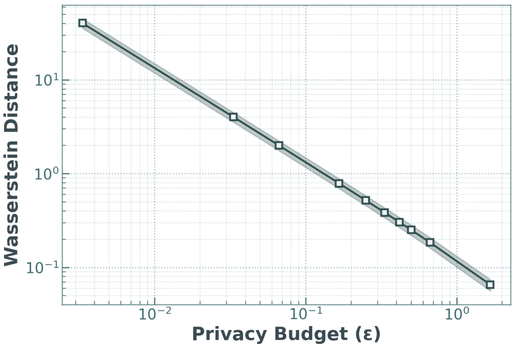

Figure 1: Tradeoff between privacy budget and change in the outcome distribution for the Jobs dataset, quantified by the Wasserstein distance between raw outcomes and the outcomes privatized through the Laplace mechanism. Smaller privacy budgets significantly alter the outcome distribution, while larger budgets preserve it but offer less privacy protection.

Following the standard practice in causal inference, we use three standard semi-synthetic datasets: IHDP (Hill, 2011), Jobs (LaLonde, 1986; Louizos et al., 2017), and LUCAS (Van der Laan & Rose, 2011). To implement these experiments, we first estimate the $\mathcal{L}_1$ sensitivity of the outcome function from the data. Using the estimated sensitivity and various privacy budgets, we determine the parameters for the Laplace noise distribution, considering the tradeoff between privacy and accuracy by selecting privacy budgets across a range of values. This noise is then used to privatize the outcomes. Details of the experimental setup are provided in Appendix E. Small budgets provide stronger privacy but distort the outcome distribution, while larger budgets offer less privacy. Figure 1 illustrates this tradeoff for the Jobs dataset, showing the change in outcome distribution—measured by the Wasserstein distance—against different budgets. We focus on moderate budget values that maintain a balance between privacy and downstream analysis validity. For ITE estimation, we compare OLS and Theil-Sen regressions over 100 trials per dataset. Figure 2 shows that Theil-Sen regression consistently reduces the error caused by the Laplace mechanism for the Jobs dataset. The observations on the other datasets yield consistent results (see Appendix E).

7 Conclusion and discussion

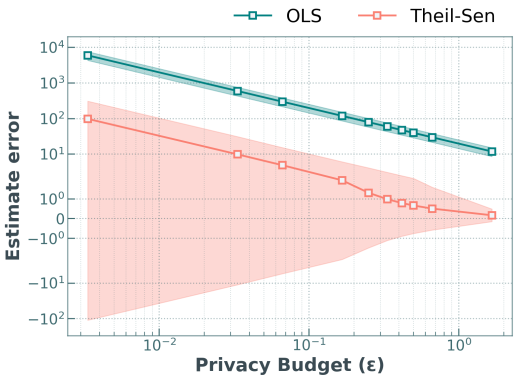

Figure 2: ITE estimation error across different privacy budgets for the Jobs dataset using Theil-Sen and OLS estimators. The solid line and the shaded region show the average and standard error over 100 trials. Theil-Sen regression consistently reduces error compared to OLS.

In this paper, we systematically investigated for the first time the impact of differential privacy (DP) on causal inference, focusing on how DP mechanisms affect the accuracy and reliability of treatment effect estimates across popular causal inference frameworks, such as RCT and DiD. Our analysis demonstrated that while applying DP to outcomes preserves the unbiasedness of ATE estimates, it can increase variance, especially under strict privacy constraints or with small sample sizes. For ITE estimates, DP can lead to significant distortions, including underestimation or invalidity of individual-level causal analysis. To address these challenges, we proposed mitigation strategies, including exact ATE recovery with privatized treatments, and robust regression to reduce ITE estimation error when outcomes are privatized. We provide a comprehensive analysis of how DP affects causal effect estimates across different settings and we propose practical strategies for mitigating DP-induced errors. These insights offer important guidance for balancing privacy and accuracy in causal inference and lay the foundation for future work in developing methods for causal representation learning from private data. Future research could extend this analysis to more complex causal frameworks or explore alternative privacy-preserving mechanisms, advancing the integration of privacy into robust causal analysis and representation learning.

Limitations. While our work provides a comprehensive analysis of differential privacy's impact on causal inference, our focus on standard global DP mechanisms may not capture the nuances introduced by other privatization methods. Additionally, since this paper primarily aims to initiate a line of research on causal inference on private data, our findings focus on standard causal inference settings. Although we propose practical mitigation strategies, they apply to the settings here studied. This work initiates research on causal representation learning from private data, pointing to the need for future methods that reduce DP-induced error. Future work could address these limitations by exploring a broader range of privacy mechanisms and causal inference scenarios, as well as proposing similar mitigation strategies tailored to a variety of causal representation learning methods.

Ethical considerations. Our work sits at the intersection of privacy and causal inference, both with significant ethical implications. Ensuring the privacy of individuals, particularly in sensitive domains such as healthcare and social sciences, is paroamount. Differential privacy provides a robust framework for this purpose, but its application must be carefully balanced against the potential impacts on the accuracy and validity of the findings, which could have serious ethical ramifications, especially in policy-making or clinical decisions. We encourage further exploration of these ethical dimensions, particularly in applying our proposed strategies, to ensure that privacy-preserving techniques do not inadvertently harm the very individuals they are designed to protect.

References

Alberto Abadie and Guido W Imbens. Matching on the estimated propensity score. Econometrica, 84(2):781–807, 2016.

Anish Agarwal and Rahul Singh. Causal inference with corrupted data: Measurement error, missing values, discretization, and differential privacy. arXiv preprint arXiv:2107.02780, 2021.

Michael G Akritas, Susan A Murphy, and Michael P Lavalley. The theil-sen estimator with doubly censored data and applications to astronomy. Journal of the American Statistical Association, 90 (429):170–177, 1995.

Daniel Alabi and Salil Vadhan. Hypothesis testing for differentially private linear regression. In Advances in Neural Information Processing Systems, volume 35, pp. 14196–14209, 2022.

Joshua D Angrist and Jörn-Steffen Pischke. Mostly Harmless Econometrics: An Empiricist's Companion. Princeton University Press, 2009.

Orley C Ashenfelter and David Card. Using the longitudinal structure of earnings to estimate the effect of training programs, 1984.

Vinod K Chauhan, Soheila Molaei, Marzia Hoque Tania, Anshul Thakur, Tingting Zhu, and David A Clifton. Adversarial de-confounding in individualised treatment effects estimation. In International Conference on Artificial Intelligence and Statistics, pp. 837–849. PMLR, 2023.

Cynthia Dwork and Aaron Roth. Algorithmic Foundations of Differential Privacy. Foundations and Trends in Theoretical Computer Science, 2014.

Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In Theory of Cryptography Conference, pp. 265–284. Springer, 2006.

Martin A Fischler and Robert C Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. Communications of the ACM, 24 (6):381–395, 1981.

Ronald A. Fisher. The Design of Experiments. Oliver and Boyd, 1935.

Xing Sam Gu and Paul R Rosenbaum. Comparison of multivariate matching methods: Structures, distances, and algorithms. Journal of Computational and Graphical Statistics, 2(4):405–420, 1993.

Sharmistha Guha and Jerome P Reiter. Differentially private estimation of weighted average treatment effects for binary outcomes. arXiv preprint arXiv:2408.14766, 2024.

Jennifer L Hill. Bayesian nonparametric modeling for causal inference. Journal of Computational and Graphical Statistics, 20(1):217–240, 2011.

Daniel E Ho, Kosuke Imai, Gary King, and Elizabeth A Stuart. Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. Political Analysis, 15(3): 199–236, 2007.

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

Guido W Imbens and Donald B Rubin. Causal Inference in Statistics, Social, and Biomedical Sciences. Cambridge University Press, 2015.

Adel Javanmard, Vahab Mirrokni, and Jean Pouget-Abadie. Causal inference with differentially private (clustered) outcomes. arXiv preprint arXiv:2308.00957, 2023.

Nathan Kallus. Generalized optimal matching methods for causal inference. Journal of Machine Learning Research, 21(62):1–54, 2020.

Vishesh Karwa and Salil Vadhan. Finite sample differentially private confidence intervals. arXiv preprint arXiv:1711.03908, 2017.

Sören R Künzel, Jasjeet S Sekhon, Peter J Bickel, and Bin Yu. Metalearners for estimating heterogeneous treatment effects using machine learning. Proceedings of the National Academy of Sciences, 116(10):4156–4165, 2019.

Matt J Kusner, Yu Sun, Karthik Sridharan, and Kilian Q Weinberger. Private causal inference. In International Conference on Artificial Intelligence and Statistics, pp. 1308–1317. PMLR, 2016.

Robert J LaLonde. Evaluating the econometric evaluations of training programs with experimental data. The American Economic Review, pp. 604–620, 1986.

Christos Louizos, Uri Shalit, Joris M Mooij, David Sontag, Richard Zemel, and Max Welling. Causal effect inference with deep latent-variable models. In Advances in Neural Information Processing Systems, volume 30, 2017.

Fengshi Niu, Harsha Nori, Brian Quistorff, Rich Caruana, Donald Ngwe, and Aadharsh Kannan. Differentially private estimation of heterogeneous causal effects. In Conference on Causal Learning and Reasoning, pp. 618–633. PMLR, 2022.

Yuki Ohnishi and Jordan Awan. Locally private causal inference for randomized experiments. arXiv preprint arXiv:2301.01616, 2023.

Judea Pearl. Causality. Cambridge University Press, 2009.

Saeyoung Rho, Rachel Cummings, and Vishal Misra. Differentially private synthetic control. International Conference on Artificial Intelligence and Statistics, pp. 1457–1491, 2023.

Peter J Rousseeuw and Annick M Leroy. Robust Regression and Outlier Detection. John Wiley & Sons, 2005.

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

Bernhard Schölkopf. Causality for machine learning. In Probabilistic and Causal Inference: The Works of Judea Pearl, pp. 765–804. Association for Computing Machinery, 2022.

Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. Toward causal representation learning. Proceedings of the IEEE, 109(5):612–634, 2021.

Pranab Kumar Sen. Estimates of the regression coefficient based on kendall's tau. Journal of the American Statistical Association, 63(324):1379–1389, 1968.

Uri Shalit, Fredrik D Johansson, and David Sontag. Estimating individual treatment effect: Generalization bounds and algorithms. In Proceedings of the 34th International Conference on Machine Learning, pp. 3076–3085. PMLR, 2017.

Or Sheffet. Differentially private ordinary least squares. In Proceedings of the 34th International Conference on Machine Learning, pp. 3105–3114. PMLR, 2017.

Michael T. Smith et al. Differentially private regression with gaussian processes. In International Conference on Artificial Intelligence and Statistics, pp. 1195–1203. PMLR, 2018.

Elizabeth A Stuart. Matching methods for causal inference: A review and a look forward. Statistical Science, 25(1):1, 2010.

Henri Theil. A rank-invariant method of linear and polynomial regression analysis. Indagationes Mathematicae, 12(85):173, 1950.

Mark J Van der Laan and Sherri Rose. Targeted Learning: Causal Inference for Observational and Experimental Data, volume 4. Springer, 2011.

Victor Veitch, Yixin Wang, and David Blei. Using embeddings to correct for unobserved confounding in networks. In Advances in Neural Information Processing Systems, volume 32, 2019.

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

Appendix / supplemental material

A Related work

In this work, we examine the impact of differential privacy (DP) mechanisms on causal inference from privatized data across various standard causal inference settings. Although relevant work on DP and causal inference is referenced throughout the paper, we now provide a more comprehensive review of related literature on these topics here.

Privacy and statistical inference. Differential privacy (DP) has become a crucial tool for ensuring that individual data points cannot be inferred from the released outcomes of statistical analyses. Foundational work by Dwork et al. (2006) and Dwork & Roth (2014) laid the groundwork for applying DP across various statistical tasks. Studies such as Karwa & Vadhan (2017); Sheffet (2017) have developed DP techniques for linear regression, while others have used Gaussian Processes for differentially private regression (Smith et al., 2018). Meanwhile, the challenge of preserving statistical validity under DP has been a critical focus, particularly in hypothesis testing for linear regression (Alabi & Vadhan, 2022).

Causal inference and differential privacy. The intersection of DP and causal inference has attracted interest, primarily focusing on developing privacy-preserving causal inference mechanisms. For example, Kusner et al. (2016) proposed methods for DP within the Additive Noise Model framework, Guha & Reiter (2024) developed algorithms for estimating weighted average treatment effects under DP, and Niu et al. (2022) introduced meta-algorithms for conditional average treatment effect estimation with DP guarantees. However, these works generally assume access to true data, leaving the effects of DP on downstream causal inference tasks underexplored. Few studies focus on causal inference from already-privatized data; Javanmard et al. (2023) propose a mechanism that improves DP-induced variance in causal estimates, and Agarwal & Singh (2021) introduces a pipeline for causal inference from corrupted data with low-rank covariates.

Our contribution. Our work addresses the critical gap in understanding how DP mechanisms influence causal estimates when applied to already-privatized data. Analyzing various standard causal inference settings and privacy mechanisms, we provide a more comprehensive assessment of the trade-offs between data privacy and the accuracy of causal estimates and offer practical guidelines for balancing these trade-offs.

B Causal inference with private outcome

In this appendix, we provide a more detailed analysis of the impact of differential privacy on causal inference when the outcome variable is privatized, following the same notation introduced in Section 4 of the main paper. For completeness, we restate the key results presented in the main paper and expand on them with additional discussion. Here we cover both ATE and ITE estimation in randomized controlled trials and difference-in-differences frameworks under private outcomes, elaborating on the implications of privacy mechanisms such as the Laplace and Gaussian mechanisms.

B.1 ATE with private outcome

Randomized controlled trial. When the data is collected through an RCT, the treatment assignment T is assumed to be independent of the covariates X, hence τ is simply estimated by the difference between expected outcome in the treatment and control groups, given in Equation 8.

$$\tau = \mathbb{E}\left[Y|T=1\right] - \mathbb{E}\left[Y|T=0\right]. \tag{8}$$

Similarly, we can compute $\tilde{\tau}$ as the difference between the conditional expectations of $\tilde{Y}$. Given an ϵ-differntially private outcome $\tilde{Y} = Y + \text{Lap}(\lambda)$, where $\lambda = \frac{\Delta f}{\epsilon}$ for both treatment and control groups, we show Proposition 4.1. This is due to the 0-mean noise added to Y to obtain $\tilde{Y}$. The detailed derivations for Proposition 4.1 are included in Appendix C. Using the Gaussian mechanism for privatizing the outcome leads to similar results, where the statement of Proposition 4.1 holds with the only difference being $\text{Var}(\bar{\tilde{\tau}}) = \text{Var}(\bar{\tau}) + \sigma^2 \left(\frac{1}{n_1} + \frac{1}{n_0}\right)$, where σ^2 is the variance of the Gaussian noise.

Difference in differences. Consider now the case of an observational study where RCT data is not available, but the DiD approach can be applied. DiD assumes parallel trends between treated and control groups–that is, in the absence of treatment, the average outcomes for the two groups would have followed the same trajectory over time (Angrist & Pischke, 2009). The key assumption here is that any differences between the groups are time-invariant and the treatment effect can be estimated as the difference in the changes in outcomes between the two groups before and after the treatment. In this case, the aggregate causal effect is quantified by the average treatment effect on the treated (ATT). Specifically, let Y_t and Y_{t-1} represent the outcomes at times t (post-treatment) and $t-1$ (pre-treatment), respectively. Then, the ATT under the DiD setup can be estimated by

$$\begin{aligned}\tau = &\left(\mathbb{E}[Y_t \mid T=1] - \mathbb{E}[Y_{t-1} \mid T=1]\right) \\ &- \left(\mathbb{E}[Y_t \mid T=0] - \mathbb{E}[Y_{t-1} \mid T=0]\right),\end{aligned} \tag{9}$$

where, for simplicity of notation, we continue using τ (and $\tilde{\tau}$) to denote the ATT in DiD. Given the expectations in equations 9 and 12, we can deduce a similar fact as what we derived for ATE in the RCT setting.

Proposition B.1. *Given a sample with $n_{1,t}$ and $n_{1,t-1}$ individuals in the treatment, and $n_{0,t}$ and $n_{0,t-1}$ individuals in the control groups, before and after intervention (respectively), we have $\tau = \tilde{\tau}$ and Var$(\bar{\tilde{\tau}})$ = Var$(\bar{\tau}) + 2\lambda^2 \left(\frac{1}{n_{1,t}} + \frac{1}{n_{1,t-1}} + \frac{1}{n_{0,t}} + \frac{1}{n_{0,t-1}}\right)$.*

While similar to Proposition 4.1, we include the calculations for Proposition B.1 in Appendix C for completeness. As we observed in the RCT case, replacing the Laplace mechanism with the Gaussian mechanism to obtain $\tilde{Y}$ via a Gaussian noise with variance σ^2 only impacts the sample variance in the statement of Proposition B.1, changing the factor of $2\lambda^2$ to σ^2.

B.2 ITE with private outcome

In the absence of averaging, the additive noise in $\tilde{Y}$ does not evaluate to 0 for individual causal estimates. Here we derive the distribution of Z_i in each of RCT and DiD cases. Due to the fundamental problem of causal inference, ITE cannot be directly computed from data. However, given the experimental design and under the assumptions of the causal inference setup, a variety of methods could be used to estimate ITE via inferring potential outcomes (see Imbens & Rubin (2015); Shalit et al. (2017); Wager & Athey (2018); Künzel et al. (2019); Chauhan et al. (2023)) or matching (see (Ho et al., 2007; Stuart, 2010; Kallus, 2020)). Here we aim to provide an analysis that could be extended to various ITE estimation methods. To this end, we consider the potential outcomes $Y_i \mid do(T=1)$ and $Y_i \mid do(T=0)$, and private potential outcomes $\tilde{Y_i} \mid do(T=1)$ and $\tilde{Y_i} \mid do(T=0)$, each privatized via an independent additive noise. We further assume that both treatment and control groups are privatized using the same noise distribution parameters. In appendix C, we extend this analysis to the more general case of different noise parameters for the two groups, which is particularly relevant in situations where the sensitivity of the outcome function is dependent on the treatment assignment.

Randomized controlled trial. Considering data from RCT with private $\tilde{Y} = Y + \text{Lap}(\lambda)$, we can write

$$\tilde{\tau}_i = \tau_i + \xi_1 - \xi_0, \tag{10}$$

where $\xi_1 \sim \text{Lap}(\lambda)$ and $\xi_2 \sim \text{Lap}(\lambda)$, and hence, the DP-induced error in ITE estimation is distributed as the distribution of $\xi_1 - \xi_0$. Computing this distribution (Appendix C), we conclude Proposition 4.2. Since the ITE error follows the distribution of $\xi_1 - \xi_0$, in the case of the Gaussian mechanism, we get $Z_i \sim \mathcal{N}\left(0, 2\sigma^2\right)$, where σ^2 is the variance of the additive Gaussian noise.

Difference in differences. With a DiD approach in observational settings, the ITE computed from $\tilde{Y} = Y + \text{Lap}(\lambda)$ can be written as

$$\tilde{\tau}_i = \tau_i + (\xi_{1,t} - \xi_{1,t-1}) + (\xi_{0,t} - \xi_{0,t-1}), \tag{11}$$

where all terms following τ_i on the right hand side of Equation 11 are Lap(λ)-distributed. Therefore, the ITE estimation error in this case is essentially distributed as the difference between two ITE errors in the RCT case, which brings us to Proposition 4.3. We prove this proposition in Appendix C, where we also state the distribution of the error. Note that using the Gaussian mechanism will have a similar implication as in the RCT case, but DiD involves four Gaussian noises, resulting in $Z_i \sim \mathcal{N}\left(0, 4\sigma^2\right)$.

C Derivation of the results for causal inference on private outcome

In this appendix we include the detailed derivation of the results on causal estimates computed from private outcomes, discussed in Section 4 and Appendix B. These derivations use the notation described in Section 4, and unless otherwise stated, we make the same assumptions.

Estimating the ATE with the Laplace mechanism in RCT. Recall that $\epsilon-$DP outcomes are defined as $\tilde{Y} = Y + \text{Lap}(\lambda)$, where $\lambda = \frac{\Delta f}{\epsilon}$. To estimate the private ATE using $\tilde{Y}$, we write

$$\tilde{\tau} = \mathbb{E}[\tilde{Y}|T=1] - \mathbb{E}[\tilde{Y}|T=0].$$

Substituting $\tilde{Y}_i$ into the expression, we obtain

$$\begin{aligned}\tilde{\tau} &= \mathbb{E}[Y|T=1] + \mathbb{E}[\text{Lap}(\lambda)|T=1] - (\mathbb{E}[Y|T=0] + \mathbb{E}[\text{Lap}(\lambda)|T=0]) \\ &= \tau + \mathbb{E}[\text{Lap}(\lambda)|T=1] - \mathbb{E}[\text{Lap}(\lambda)|T=0].\end{aligned}$$

Since the noise added by the Laplace mechanism is independent of the treatment assignment and has zero mean, the expected value of the Laplace noise is 0, and hence $\mu = \tilde{\mu}$ as stated in Proposition 4.1. Following similar steps, we can write the sample variance of the ATE estimated from private outcomes as

$$\text{Var}(\bar{\tilde{\tau}}) = \frac{1}{n_1}\left[\text{Var}(Y|T=1) + \text{Var}(\text{Lap}(\lambda))\right] + \frac{1}{n_0}\left[\text{Var}(Y|T=0) + \text{Var}(\text{Lap}(\lambda))\right],$$

where the Var(Lap(λ)) terms are due to the Laplace noise added to the outcomes for the treatment and control groups. Using the facts that $\text{Var}(\bar{\tau}) = \frac{1}{n_1}\text{Var}(Y|T=1) + \frac{1}{n_0}\text{Var}(Y|T=0)$ and $\text{Var}(\text{Lap}(\lambda)) = 2\lambda^2$, we obtain

$$\text{Var}(\bar{\tilde{\tau}}) = \text{Var}(\bar{\tau}) + 2\lambda^2\left(\frac{1}{n_1} + \frac{1}{n_0}\right),$$

which completes the derivation of the results stated in Proposition 4.1. As explained in Section 4.1, using the Gaussian mechanism simply replaces the $2\lambda^2$ factor in the equation above with the variance of the Gaussian noise.

Estimating the ATT with the Laplace mechanism in DiD. The impact of privatizing Y using a Laplace noise on the average treatment effect on the treated (ATT) in a DiD setting, is similar to its impact on ATE in an RCT setting. This is due to the fact that the ATT estimator in DiD can be formulated as the difference between two RCT estimators at time steps t and $t-1$. Specifically, for DiD we can write

$$\begin{aligned}\tilde{\tau} &= \Big(\mathbb{E}[\tilde{Y}_t \mid T=1] - \mathbb{E}[\tilde{Y}_{t-1} \mid T=1]\Big) - \Big(\mathbb{E}[\tilde{Y}_t \mid T=0] - \mathbb{E}[\tilde{Y}_{t-1} \mid T=0]\Big), \qquad (12) \\ &= \Big(\mathbb{E}[\tilde{Y}_t \mid T=1] - \mathbb{E}[\tilde{Y}_t \mid T=0]\Big) - \Big(\mathbb{E}[\tilde{Y}_{t-1} \mid T=1] - \mathbb{E}[\tilde{Y}_{t-1} \mid T=0]\Big).\end{aligned}$$

It follows immediately that the ATE estimate remains unbiased as in the RCT case, and with the sample sizes $n_{1,t}$ and $n_{1,t-1}$ of the treatment, and $n_{0,t}$ and $n_{0,t-1}$ of the control groups, before and after intervention, the variance becomes

$$\text{Var}(\bar{\tilde{\tau}}) = \text{Var}(\bar{\tau}) + 2\lambda^2\left(\frac{1}{n_{1,t}} + \frac{1}{n_{1,t-1}} + \frac{1}{n_{0,t}} + \frac{1}{n_{0,t-1}}\right),$$

as stated in Proposition B.1. Similar to the RCT case, replacing the Laplace mechanism with the Gaussian mechanism results in replacing $2\lambda^2$ above with the variance of the Gaussian noise.

Estimating the ITE with the Laplace mechanism in RCT. Following the discussion in Section 4.2, we derive the ITE results using the potential outcomes. Given $\epsilon-$DP outcome $\tilde{Y} = Y + \text{Lap}(\lambda)$, we can write the ITE estimated from $\tilde{Y}$ in an RCT setting as

$$\tilde{\tau}_i = \tau_i + \xi_1 - \xi_0,$$

where $\xi_1 \sim \text{Lap}(\lambda_1)$ and $\xi_0 \sim \text{Lap}(\lambda_0)$. While in Section 4.2 we considered the case where $\lambda_1 = \lambda_0 = \lambda$, here we allow the privacy parameters for the treatment and control to be different [1], and we derive the results for both cases –where $\lambda_1 = \lambda_0 = \lambda$ and where $\lambda_1 \neq \lambda_0$. Hence, to find the distribution of $Z_i = \tilde{\tau}_i - \tau_i$, we calculate the probability density function (PDF) of $\xi_1 - \xi_0$ as

$$\begin{aligned} f_{Z_i}(z) = &\int_{-\infty}^{+\infty} f_{\xi_1}(u) f_{\xi_0}(u-z) du \\ =& \frac{1}{4\lambda_1\lambda_0} \int_{-\infty}^{+\infty} \exp\left(-\frac{|u|}{\lambda_1}\right) \exp\left(-\frac{|u-z|}{\lambda_0}\right) du \\ =& \frac{1}{4\lambda_1\lambda_0} \int_{-\infty}^{0} \exp\left(\frac{u}{\lambda_1}\right) \exp\left(\frac{u-z}{\lambda_0}\right) du \\ &+ \frac{1}{4\lambda_1\lambda_0} \int_{0}^{z} \exp\left(-\frac{u}{\lambda_1}\right) \exp\left(\frac{u-z}{\lambda_0}\right) du \\ &+ \frac{1}{4\lambda_1\lambda_0} \int_{z}^{+\infty} \exp\left(-\frac{u}{\lambda_1}\right) \exp\left(-\frac{u-z}{\lambda_0}\right) du, \end{aligned}$$

where, in the third line we used the assumption that $z \geq 0$, which, as we discuss in Section 4, we impose for simplicity, though extension to $z < 0$ is straightforward. Computing the integrals and simplifying, we obtain

$$f_{Z_i}(z) = \begin{cases} \frac{1}{2\lambda_1^2 - \lambda_2^2} \left[\lambda_1 \exp\left(-\frac{z}{\lambda_1}\right) - \lambda_0 \exp\left(-\frac{z}{\lambda_0}\right)\right] & \lambda_1 \neq \lambda_0 \\ \frac{1}{4\lambda} \exp\left(-\frac{z}{\lambda}\right) \left[1 + \frac{z}{\lambda}\right] & \lambda_1 = \lambda_0 = \lambda \end{cases},$$

which completes the derivation of the error distribution stated in Proposition 4.2, as well as the more general case where $\lambda_1 \neq \lambda_0$. Using this distribution, it is straightforward to calculate the expected error as

$$\mathbb{E}\left[Z_i\right] = \int_0^{+\infty} z f_{Z_i}(z) dz = \int_0^{+\infty} \frac{z}{4\lambda} \exp\left(-\frac{z}{\lambda}\right) \left[1 + \frac{z}{\lambda}\right] dz = \frac{3}{2}\lambda^2.$$

Estimating the ITE with the Laplace mechanism in DiD. Similar to the RCT case, the error in ITE estimation with $\tilde{Y} = Y + \text{Lap}(\lambda)$, can be written as

$$\begin{aligned} z_i &= \tilde{\tau}_i - \tau_i \\ &= \left(\xi_{1,t} - \xi_{1,t-1}\right) - \left(\xi_{0,t} - \xi_{0,t-1}\right), \end{aligned} \tag{13}$$

where, now we have two pairs of differences between Laplace-distributed variables. To derive the results in a more general case than Proposition B.1, here we allow the treated and control groups to have different Laplace parameters, where $\xi_{1,t} \sim \text{Lap}(\lambda_1)$, $\xi_{1,t-1} \sim \text{Lap}(\lambda_1)$, $\xi_{0,t} \sim \text{Lap}(\lambda_0)$, and $\xi_{0,t-1} \sim \text{Lap}(\lambda_0)$. Note that the results for the case where different time steps, rather than treatment groups, have different Laplace parameters yields identical results, since we can rearrange Equation 13 to write the error as $(\xi_{1,t} - \xi_{0,t}) - (\xi_{1,t-1} - \xi_{0,t-1})$, which is again the difference between two pairs of Laplace noises with the same parameters. While deriving the error distribution in the RCT case, we already computed the distributions of the terms in each bracket on the right hand side of Equation 13:

$$\begin{aligned} \left(\xi_{1,t} - \xi_{1,t-1}\right) &\sim \frac{1}{4\lambda_1} \exp\left(-\frac{z}{\lambda_1}\right) \left[1 + \frac{z}{\lambda_1}\right] \\ \left(\xi_{0,t} - \xi_{0,t-1}\right) &\sim \frac{1}{4\lambda_0} \exp\left(-\frac{z}{\lambda_0}\right) \left[1 + \frac{z}{\lambda_0}\right]. \end{aligned}$$

[1] The $\lambda_1 \neq \lambda_0$ case is particularly relevant when the outcome function has different sensitivities for the treatment and control groups.

Hence, following similar steps as in the RCT case, we can calculate the PDF of Z_i as

$$\begin{aligned} f_{Z_i}(z) =& \int_0^{+\infty} f_{\xi_{1,t}-\xi_{1,t-1}}(u) f_{\xi_{0,t}-\xi_{0,t-1}}(u-z) du \\ =& \frac{1}{16\lambda_1\lambda_0} \int_0^{+\infty} \exp\left(\frac{-u}{\lambda_1} + \frac{z-u}{\lambda_0}\right) \left[1 + \frac{z}{\lambda_1}\right] \left[1 - \frac{z-u}{\lambda_0}\right] du, \end{aligned}$$

where, for simplicity, we considered the case of non-negative pairs of differences, while extension to the entire real-line is straightforward. Computing the integral above, we obtain the following error distribution.

$$f_{Z_i}(z) = \frac{e^{\frac{z}{\lambda_0}}}{16\lambda_1\lambda_0} \left[1 + \frac{z}{\lambda_1}\right] \left[1 - \frac{z}{\lambda_0} + \frac{\lambda_1}{\lambda_1 + \lambda_0}\right],$$

which becomes $f_{Z_i}(z) = \frac{\lambda+z}{32\lambda^2} e^{\frac{z}{\lambda}} \left(\frac{3}{2} - \frac{z}{\lambda}\right)$ when $\lambda_1 = \lambda_0 = \lambda$. Calculating the expectation of this distribution, in either of $\lambda_1 = \lambda_0 = \lambda$ or $\lambda_1 \neq \lambda_0$ cases, yields a divergent integral, which means the mean of the error distribution in this case is infinite. This completes the proof for Proposition 4.3.

D Derivation of the results for causal inference on private treatment

In this appendix we include the detailed derivation of the results on causal estimates computed from private treatments, discussed in Section 5. Here we use the notation described in Section 5, and unless otherwise stated, we make the same assumptions.

Estimating the ATE with private treatments. Recall that the binary treatment T is privatized using the randomized response mechanism, where the private treatment $\tilde{T}$ is kept equal to T with probability p and flipped to $1 - T$ with probability $1 - p$. Therefore, for a real-valued outcome Y, we can write $\tilde{\mu}_1 \equiv \mathbb{E}\left[Y \mid \tilde{T} = 1\right]$ as

$$\tilde{\mu}_1 = \mathbb{P}\left(T = 1 \mid \tilde{T} = 1\right) \mu_1 + \mathbb{P}\left(T = 0 \mid \tilde{T} = 1\right) \mu_0, \tag{14}$$

by the law of total expectation and using the fact that $Y \perp\!\!\!\perp \tilde{T} | T$, which holds since we assumed a structural causal model $Y = f(X, T)$, and $\tilde{T}$ is randomized based on T by the parameter p, independent of the other variables. Similarly, we can write $\tilde{\mu}_0 \equiv \mathbb{E}\left[Y \mid \tilde{T} = 0\right]$ as

$$\tilde{\mu}_0 = \mathbb{P}\left(T = 1 \mid \tilde{T} = 0\right) \mu_1 + \mathbb{P}\left(T = 0 \mid \tilde{T} = 0\right) \mu_0. \tag{15}$$

Using the Baye's rule, we can compute the probabilities in equations 14 and 15 as

$$\begin{aligned} \mathbb{P}\left(T = 1 \mid \tilde{T} = 1\right) &= \frac{pq}{pq + (1-p)(1-q)} & \mathbb{P}\left(T = 0 \mid \tilde{T} = 1\right) &= \frac{(1-p)(1-q)}{pq + (1-p)(1-q)} \\ \mathbb{P}\left(T = 1 \mid \tilde{T} = 0\right) &= \frac{(1-p)q}{pq + (1-p)(1-q)} & \mathbb{P}\left(T = 0 \mid \tilde{T} = 0\right) &= \frac{p(1-q)}{pq + (1-p)(1-q)}, \end{aligned}$$

Given these values and the equations 14 and 15, we can write the system of linear equations for μ_1 and μ_0 as

$$\begin{bmatrix} \frac{pq}{pq+p'q'} & \frac{p'q'}{pq+p'q'} \\ \frac{p'q}{pq+p'q'} & \frac{pq'}{pq+p'q'} \end{bmatrix} \begin{bmatrix} \mu_1 \\ \mu_0 \end{bmatrix} = \begin{bmatrix} \tilde{\mu}_1 \\ \tilde{\mu}_0 \end{bmatrix}, \tag{16}$$

where, for ease of notation, we used $p' := 1 - p$ and $q' := 1 - q$. We can solve Equation 16 for μ_1 and μ_0, which yields

$$\mu_1 = \frac{\tilde{\mu}_0 - \frac{\beta_0}{\beta_1}\tilde{\mu}_1}{\alpha_0 - \frac{\beta_0}{\beta_1}\alpha_1} \qquad \mu_0 = \frac{\tilde{\mu}_1 - \alpha_1\mu_1}{\beta_1},$$

where $\alpha_1 := \frac{pq}{pq+p'q'}$, $\beta_1 := \frac{p'q'}{pq+p'q'}$, $\alpha_0 := \frac{p'q}{pq+p'q'}$, and $\beta_0 := \frac{pq'}{pq+p'q'}$, are the entries of the coefficient matrix in Equation 16. This allows us to exactly recover the true ATE in the RCT setting as $\mu_1 - \mu_0$, when $p > \frac{1}{2}$. This completes the proof for Proposition 5.1. Using these solutions, we can also compute the ATE error $Z = \tilde{\tau} - \tau$, and explore how the error changes with respect to q and p.

Estimating the ITE with private treatments. As we discuss in Section 5.2, there are several approaches for inferring ITE from data and in this paper we analyze the ITE error from privatized treatments in a matching setting. Following the notation described in Section 5.2, when matching a treated individual i to an individual j from the control group, we can write the ITE in terms of potential outcomes as $\tau_{i\sim j} = Y_i|do(T_i = 1) - Y_j|do(T_j = 0)$. Therefore, given the privatized treatments $\tilde{T}_i$ and $\tilde{T}_j$, we have

$$Y_i|do(T_i = \tilde{t}_i) - Y_j|do(T_j = \tilde{t}_j) = \begin{cases} Y_i|do(T_i = 1) - Y_j|do(T_i = 0) & \tilde{t}_i = t_i \,\&\, \tilde{t}_j = t_j \\ Y_i|do(T_i = 1) - Y_j|do(T_i = 1) & \tilde{t}_i = t_i \,\&\, \tilde{t}_j = 1 - t_j \\ Y_i|do(T_i = 0) - Y_j|do(T_i = 0) & \tilde{t}_i = 1 - t_i \,\&\, \tilde{t}_j = t_j \\ Y_i|do(T_i = 0) - Y_j|do(T_i = 1) & \tilde{t}_i = 1 - t_i \,\&\, \tilde{t}_j = 1 - t_j \end{cases},$$

where $\tilde{t}_i$ and $\tilde{t}_j$ are the realized values of $\tilde{T}_i$ and $\tilde{T}_j$. Using the fact that the probability of flipping each of T_i and T_j is p and T_i and T_j are privatized independently, we can write

$$\begin{aligned} \mathbb{E}\left[\tilde{\tau}_{i\sim j}\right] =& \left[Y_i(1) - Y_j(0)\right] p^2 + \left[Y_i(1) - Y_j(1)\right] p(1-p) \\ &+ \left[Y_i(0) - Y_j(0)\right] p(1-p) + \left[Y_i(0) - Y_j(1)\right] (1-p)^2 \\ =& p\left[Y_i(1) - Y_j(0)\right] + (1-p)\left[Y_i(0) - Y_j(1)\right] \\ =& p\left(\tau_i + \tau_j\right) - \tau_{j\sim i}, \end{aligned}$$

where, for ease of notation, we adopted $Y_i(t)$ to denote $Y_i|do(T_i = t)$. The derivation above gives us the expected ITE. Similarly, we can write the expected ITE error when matching i to j as

$$\begin{aligned} \mathbb{E}\left[Z_{i\sim j}\right] =& 0p^2 + p(1-p)\left[Y_j(0) - Y_j(1)\right] + p(1-p)\left[Y_i(0) - Y_i(1)\right] \\ &+ (1-p)^2\left[Y_i(0) - Y_i(1) + Y_j(0) - Y_j(1)\right] \\ =& \left(\tau_i + \tau_j\right)(1-p), \end{aligned}$$

which completes the derivation of the results in Proposition 5.2.

E Robust regression experimental setup

In this appendix, we provide additional details and observations regarding the mitigation strategy discussed in Section 6, using robust regression to reduce DP-induced errors in causal effect estimates.

Data and experimental setup. We conduct our experiment using three standard causal inference datasets featuring continuous outcomes: IHDP (Hill, 2011), Jobs (LaLonde, 1986; Louizos et al., 2017), and LUCAS (Van der Laan & Rose, 2011). Following the standard practice in causal inference, these datasets are semi-synthetic with empirically observed features paired with simulated treatments and potential outcomes (Hill, 2011; Shalit et al., 2017; Veitch et al., 2019). To quantify the impact of differential privacy on ITE estimation, we first estimate the $\mathcal{L}_1$ sensitivity of the outcome function from the data, calculating the maximum distance between datasets that differ by only a single data point. Using the estimated sensitivity and various privacy budgets, we determine the parameters for the Laplace noise distribution, which is then applied to privatize the outcomes.

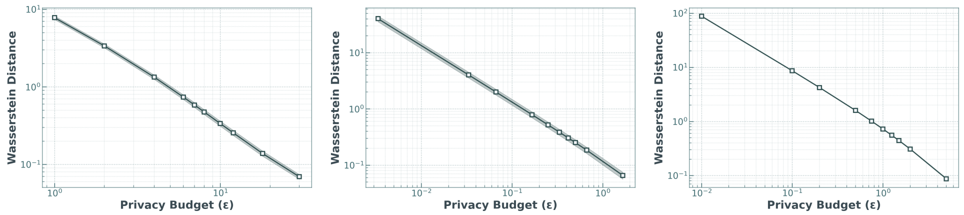

Figure 3: Tradeoff between privacy budget and change in outcomes distribution for IHDP (left), Jobs (middle), and LUCAS (right) datasets. The change in distributions is quantified by the Wasserstein distance between raw outcomes and the outcomes privatized through the Laplace mechanism. Smaller privacy budgets significantly alter the outcome distribution, while larger budgets preserve it but offer less privacy protection.

Privacy budget and distribution change. The privacy budgets are selected from a range to explore the tradeoff between privacy and accuracy. Small privacy budget values, while providing strong privacy guarantees, substantially alter the outcome distribution, rendering downstream causal analysis less reliable. Conversely, very large budgets, though preserving the original outcome distribution, offer limited privacy protection. This tradeoff is illustrated in Figure 3 for all three datasets, where we show the change in the distribution of outcomes—quantified by the Wasserstein distance—against different privacy budget values. We select more values in a moderate range of privacy budget that correspond to sufficiently small Wasserstein distances, ensuring both privacy and the validity of downstream analysis.

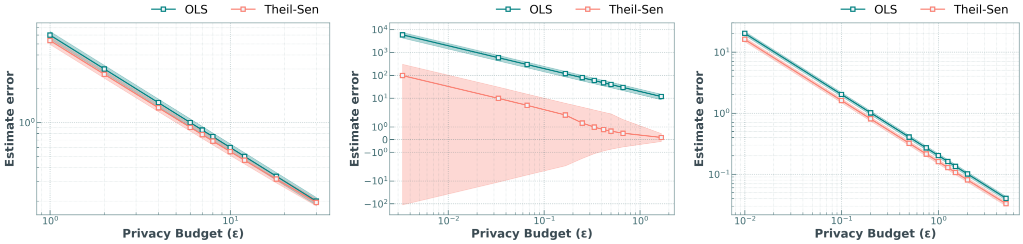

Figure 4: ITE estimation error across different privacy budgets for IHDP (left), Jobs (middle), and LUCAS (right) datasets using Theil-Sen and OLS estimators. The solid line and the shaded region show the average and standard error over 100 trials. Theil-Sen consistently reduces the error compared to OLS.

Implementation of robust regression experiments. Given the privatized outcomes, we compare the DP-induced ITE estimation error using OLS and Theil-Sen regressions. We conduct this comparison over 100 trials for each dataset. The mean and standard error of these trials, presented in Figure 4 for all three datasets, indicate that the Theil-Sen estimator consistently reduces the error the Laplace mechanism introduces to the ITE estimation across all datasets.