Improving Privacy for Respondents in Randomized Controlled Trials: A Differential Privacy Approach

Soumya Mukherjee[1], Aratrika Mustafi[1], Aleksandra Slavković[1], and Lars Vilhuber[2]

[1]Penn State, Department of Statistics
[2]Cornell University, Department of Economics

August 26, 2024

Abstract

Randomized controlled trials (RCTs) have become a powerful tool for assessing the impact of interventions and policies in many contexts. Researchers have published an increasing number of studies that rely on RCTs for at least part of the inference, and these studies typically include the response data collected, de-identified, and sometimes protected through traditional disclosure limitation methods. In our presentation and extended paper, we explore the impact of applying differentially private methods on inference, computational feasibility and accuracy, as a case study on a published article.

We find, not surprisingly, that robust but naïve methods yield strong protection, but at great loss of inference ability. However, we also explore how a partial targeted relaxation as well as a model-specific protection method can alleviate those concerns, at least in this one case study.

The case study is part of a larger research program exploring the feasibility of stronger privacy methods in real-world contexts. We briefly outline some of the consequences of applying these methods for data openness, research transparency, and privacy of respondents.

1 Introduction

Randomized controlled trials (RCTs) have become a powerful tool for assessing the impact of interventions and policies in many contexts (e.g., in economics, see the Nobel Prize lecture by Duflo, 2020). Today, they are considered the gold standard for inference in the biomedical fields and many social sciences. In economics, much of the growth has been since the 1990s. Researchers have published an increasing number of studies that rely on RCTs for at least part of the inference. In economics, the American Economic Association (AEA)'s Social Science Registry had recorded over 8,000 registrations as of 2023, with more than 12,400 researchers associated with these registrations (Vilhuber, 2024).

In a parallel development, the quest for improved transparency in the social sciences has led to more supplementary materials accompanying articles being made public as "replication packages". For instance, the flagship journal of the AEA, the American Economic Review (AER), has required analysis data and code since 2004 (Bernanke, 2004), and the AEA journals launched in 2009 implemented such a policy from creation.[1] The increased availability of complete replication packages has allowed other researchers to leverage the materials, and conduct re-analyses and meta-analyses, furthering our understanding of the methods as well as of the conclusions drawn from these studies.

The third development has been in the area of privacy protection methodology. While privacy protection has long been a standard part of the toolkit of statistical agencies (e.g., Hundepool et al. (2012)) and ethical research guidelines, new formal privacy mechanisms, such as various forms that satisfy differential privacy (DP)(Dwork et al., 2006, 2016; Dwork and Roth, 2014), offer transparent, mathematically provable, and arguably stronger promises of confidentiality.

Nevertheless, the typical (recent) guidance followed by researchers who conduct RCTs primarily relies on weaker methods, such as de-identification (Department of Health and Human Services, 2012; Kopper, Sautmann and Turitto, 2020; DIME, 2020) and other traditional

[1] See Vilhuber (2020) and Vlaeminck (2021) for a review of reproducibility practices in economics.

disclosure avoidance methods (e.g., k-anonymity (Samarati and Sweeney, 1998), l-diversity (Machanavajjhala et al., 2006), and other aggregation-based methods (Hundepool et al., 2012)). Most data included within replication packages allow for the exact reproduction of the results in the papers, and is typically simply de-identified. We are not aware of the application of DP in the context of the dissemination of data collected as part of RCTs, though some (rare) surveys are conducted using randomized response.[2] This suggests that much of the current literature on RCTs publishes replication packages that contain inadequately protected data. This is particularly concerning in the economic data setting we are exploring because many of these studies have data from respondents in low- and middle-income countries, where there may be lower legal protections than in Europe or North America.

We postulate that one of the possible reasons for the absence of strong privacy protection methods in the more recent literature is that in contrast to many of the traditional disclosure avoidance tools, the landscape for formal privacy protection tools is still evolving, and does not have tools available for straightforward efficient use by non-specialists.[3] Efficiency here is defined as "perturbing inference as little as possible compared to the unprotected inference."[4]

The present article is part of a project that aims to provide an assessment of the feasibility of using privacy enhancing technologies (PETs), in particular differentially private methods, for data publication and adjusted inference in the context of RCTs. Broadly, we aim to contribute on two separate dimensions. First, in the context of data collected for RCTs, we investigate the feasibility of preserving the generic quality of inference obtained using the confidential data even when the same inference is performed using data endowed with privacy

[2]Randomized response (Warner, 1965) is shown to be a (distributed) DP mechanism, used at the collection stage, and has been applied in various surveys, in particular on sensitive topics. A recent example explicitly referencing it in data collection within an RCT is Kancharla and Kang (2021); the earliest (central) DP work is of Vu and Slavkovic (2009).

[3]We are aware of the Two Ravens tool (D'Orazio, Deng and Shoemate, 2018), but have not seen usage of it in the space we surveyed.

[4]Inference even in the "unprotected" case is already subject to uncertainty that is often not adequately taken into account (Meager, 2019). This is even more important for the uncertainty and data modifications that are generated through statistical disclosure limitation (SDL). Abowd and Schmutte (2015) and Slavkovic and Seeman (2022) argue for the need to account for the privacy-preserving noise in analyses. Slavkovic and Seeman (2022), and references therein, discuss a general way to adjust for privacy-preserving noise in addition to other sources of uncertainty.

protections that are stronger than the simple de-identification usually used. Second, we do so while maintaining the feasibility of application, here defined as computational feasibility on commodity hardware used by researchers in these fields (and in particular, by researchers in low- and middle-income countries, where many RCTs are conducted). We do so while still striving to maintain high transparency, i.e., the ability to continue publishing datasets that allow others to verify any empirical claims made. We thus pursue three goals motivated by a setting where the typical researcher wants to do the following:

Goal 1 publish a sufficiently precise privacy-protected inference of the effect of the treatment on the treated from the model, given the data D;

Goal 2 release (publish) the privacy-protected database $\widetilde{D}$ so that others can scrutinize the analysis, while preserving the privacy of the respondents whose data are contained in the database;

Goal 3 Apply privacy-protection methods in a way that is computationally tractable on commodity computer hardware, i.e., a researcher laptop or at most a reasonably dimensioned server.

Our focus on RCTs is intentionally narrow.[5] We believe that exploring the impact of privacy-preserving technologies in the context of RCTs is useful for several reasons. First, statistical methods are, in general, straightforward: standard linear regression, difference-in-difference methods, and possibly even simple difference in means across treated and untreated populations. These are amongst the first analysis methods for which adaptations to DP protection have been studied (e.g., Awan and Slavković, 2020; Alabi et al., 2020; Slavkovic and Molinari, 2021; Barrientos et al., 2018; Bowen et al., 2020). If formal privacy-preserving methods cannot be made to work "out-of-the-box" and at scale in this context, then it will be much more difficult to argue for broader application. Second, most RCTs are small-scale, using samples of the overall population, allowing us to avoid computational constraints when

[5] A somewhat different approach is taken by Rosenblatt et al. (2023), who start with frequently-used published datasets and explore the (conceptual) reproducibility of analyses in articles that used such datasets. They, too, focus on the simpler methods and find mixed results.

algorithms scale with sample size N.[6] Third, RCTs are often accompanied by pre-analysis plans, with specific hypotheses in mind and with the intent to avoid false discovery. These areas have also been explored within the DP framework (e.g., Vu and Slavkovic, 2009; Pistner, 2020; Dwork, Su and Zhang, 2021)). Furthermore, it is already understood in the privacy community that the inherent noisiness of the sampling may affect inference (e.g., Slavkovic and Seeman, 2022). The analogy between adding noise for the purpose of BHA (Meager, 2019), and adding noise for privacy protection may be convenient to improve the acceptance of such methods. A similar Bayesian framework can be used to adjust noisy inference due to privacy (e.g., Seeman, Slavkovic and Reimherr, 2020).

2 Methods

In our full paper Mukherjee et al. (TBD), we conduct a case study, using the data and analysis code from a single published article — Blattman, Jamison and Sheridan (2017*a*) [henceforth "BJS"]. Data and code are taken from their replication package (Blattman, Jamison and Sheridan, 2017*b*). BJS report results from an RCT conducted in Liberia, in which young delinquents were offered cognitive behavioral therapy, money, or both, in order to induce them to reduce criminal and violent behavior. They measured a variety of outcomes, constructed an index that summarizes these various outcomes, and analyze the effects of the experiment. Not all interventions worked, and only a few outcomes are correlated with the intervention in a statistically significant way.[7] We apply our privacy-preserving mechanism to the data collected by BJS and then apply BJS's analysis methods to the protected data. The protected data is constructed so as to be a drop-in replacement for the data originally published by BJS. Importantly, we explicitly do not change the analysis methods used by the authors, except where necessary to take into account the privacy-preserving mechanism.[8]

[6] We note that sampling might also allow us to leverage privacy-amplifying methods (Balle, Barthe and Gaboardi, 2018), though we do not exploit that in this paper. We also note that DP was originally designed with large rather than small samples in mind but the small-scale databases are as equally important in practice.

[7] For more details, interested readers should consult the article.

[8] There are possibly many ways that other researchers might have analyzed the data collected by this (or other) sets of authors. We do not explore those, but note that it is precisely through the availability of

2.1 Basic setup

We use DP, which relies on the concept of ϵ-indistinguishability (Dwork et al., 2006; Machanavajjhala and Kifer, 2015). An algorithm M satisfies (ϵ, δ)**-differential privacy** (approximate-DP) for some $\epsilon, \delta > 0$ if for each of its possible outputs ω and for every pair of databases D_1, D_2 that differ on the addition or removal of a single record, $P(M(D_1) = \omega) \leq e^{\epsilon} P(M(D_2) = \omega) + \delta$. When $\delta = 0$, an algorithm M satisfies ϵ**-differential privacy** (pure-DP). In other words, given the outputs from two databases that differ only in a single record are very "similar", it is statistically very difficult to know if any particular record was included in the database or not. The parameter $\epsilon > 0$ is used to quantity the privacy loss. There are many algorithms and mechanisms that satisfy these and related definitions (e.g., see Desfontaines and Pejó, 2022). The fact that the outputs should be "similar" should in principle help with maintaining similarity with inferences.

Consider a de-identified dataset $D = [Y, T, X]$. For convenience, we will call the de-identified dataset "confidential," because our goal is to truly anonymize it, recognizing that de-identification is usually insufficient. The regression model of interest in the absence of blocking variables is given by

$$Y = T\tau + X\gamma + e, \tag{1}$$

where T is $n \times t$ matrix of exhaustive assignment to treatment arms, τ is t-dimensional vector of treatment effects, X is $n \times p$ matrix of explanatory covariates, γ is a p-dimensional vector of coefficients, and e is n-dimensional error term with $e_i \overset{i.i.d}{\sim} N(0, \sigma^2)$. The typical parameter of interest is the treatment effect $\hat{\tau}$, estimated from (1), with regression coefficients $\hat{\gamma}$ often ignored or of lesser importance.

2.2 Algorithms

We propose three different but related algorithms for privacy-protecting the confidential data D and the inference concerning the treatment effects τ. All of them rely on a common base

replication packages that such differing approaches can be addressed – through replications.

algorithm, applied to the covariate data X and treatment assignments T, but differ in how they treat outcome variables Y and treatment effects τ. In all cases, our goal is to release a sanitized replication database $[\widetilde{Y}, \widetilde{T}, \widetilde{X}]$ together with a sanitized inference concerning the treatment effects τ (which includes treatment effect point and interval estimates, standard errors and p-values for tests of significance of treatment effects). The algorithms differ in how they process Y and associated inferences. We summarize the algorithms here, for details, in particular on how to compare privacy loss across all three algorithms, see Mukherjee et al. (TBD).[9] Mukherjee et al. (2023) describes Algorithm 1 as presented at the 2023 NBER workshop.

2.3 Base Algorithm

To create ϵ-DP privacy-protected covariate data $\widetilde{X}$, we sample from an appropriately perturbed multivariate histogram of X (e.g., see Dwork et al., 2006; Wasserman and Zhou, 2010). The histogram counts are sanitized using the Laplace mechanism. Continuous variables are discretized into bins using 'precision' ζ before we create the histogram.[10] We sample from the protected histogram to generate ϵ-DP protected $\widetilde{X}$. We then reimplement the treatment design, by doing random assignment for n synthetic treatment units,[11] thus creating ϵ-DP protected treatment indicators $\widetilde{T}$.

2.4 Algorithm 1: Hybrid-DP

Given $\left[\widetilde{X}, \widetilde{T}\right]$, the protected response $\widetilde{Y}$ is created using a generative model that depends on standard statistical inferential procedures performed on the confidential data D. The generative model can use $\hat{\tau}$ obtained from estimating (1) on $[X, T]$, but could also use different generative models (we report on such experiments in the full paper). $\widetilde{Y}$ is obtained as

[9] The notation here is simplified for the sake of discussion.

[10] The precision of binning prescribes how many bins are created for each continuous variable dimension, as a function of the sample size (for e.g., if $\zeta = \frac{2}{3}$ and $p = 2$, with both variables being continuous, the total number of bins in the multivariate histogram is $\left(n^{\frac{2}{3}}\right)^2 = n^{\frac{4}{3}}$).

[11] In practice, we set n to be the same number as in the original X, though this is not strictly necessary.

$$\widetilde{Y} := \widetilde{T}\hat{\tau} + \widetilde{X}\hat{\gamma} + E, \tag{2}$$

where $E \sim N(0, \hat{\sigma}^2 I)$. We then estimate (1), replacing $[Y, T, X]$ with $[\widetilde{Y}, \widetilde{T}, \widetilde{X}]$, to obtain the releaseable $\widetilde{\tau}$. The confidential $\hat{\tau}$ are not published.

By using the parameter estimates $[\hat{\tau}, \hat{\gamma}]$ obtained from the confidential data D in this algorithm, it is expected that inference validity is preserved by the protection procedure. The released $\widetilde{\tau}$ is nevertheless perturbed and therefore conventionally protected because it is based on ϵ-DP-protected $\widetilde{X}$ and model-based imputation, but is not formally DP-protected. The privacy-protected $\widetilde{D}_1(\epsilon, \zeta) = \left[\widetilde{Y}, \widetilde{T}, \widetilde{X}\right]$ is published as part of the replication package, and $\widetilde{\tau}$, its associated standard errors and inferential statistics based on (1), are estimated using $\widetilde{D}$, and published as part of the relevant article.

2.5 Fully DP algorithms

In Mukherjee et al. (TBD), we develop two additional algorithms. Algorithm 2 develops a model-agnostic fully DP algorithm that simply provides protection to the underlying data. As in Algorithm 1, all data can be released, and the expectation is that the published $\hat{\tau}$ is derived from the released data. The data thus released, $\widetilde{D}_2(\epsilon, \zeta) = \left[\widetilde{Y}_2, \widetilde{T}_2, \widetilde{X}_2\right]$, since not constrained or derived from a particular model, should have wider utility, but may not provide enough utility for the specific purpose of the article. In the paper, we experiment with a variety of ways to generate $\widetilde{Y}$, both independently of $[T, X]$ and jointly with $[T, X]$.

We also propose Algorithm 3 in an attempt to improve the accuracy of inferences, while retaining full DP protection, by again relying on model-based mechanisms. We rely on methods first proposed by Karwa and Vadhan (2017) and Kazan et al. (2023) to compute DP-protected point estimates, standard errors, 95% confidence intervals and p-values corresponding to tests of significance for τ. The computation of such privacy-protected parameters still requires access to the confidential data, and thus cannot be reproduced by readers of the article. Data releases can either be the same (independently produced) $\widetilde{D}_2(\epsilon, \zeta)$ from Algorithm 2, or a

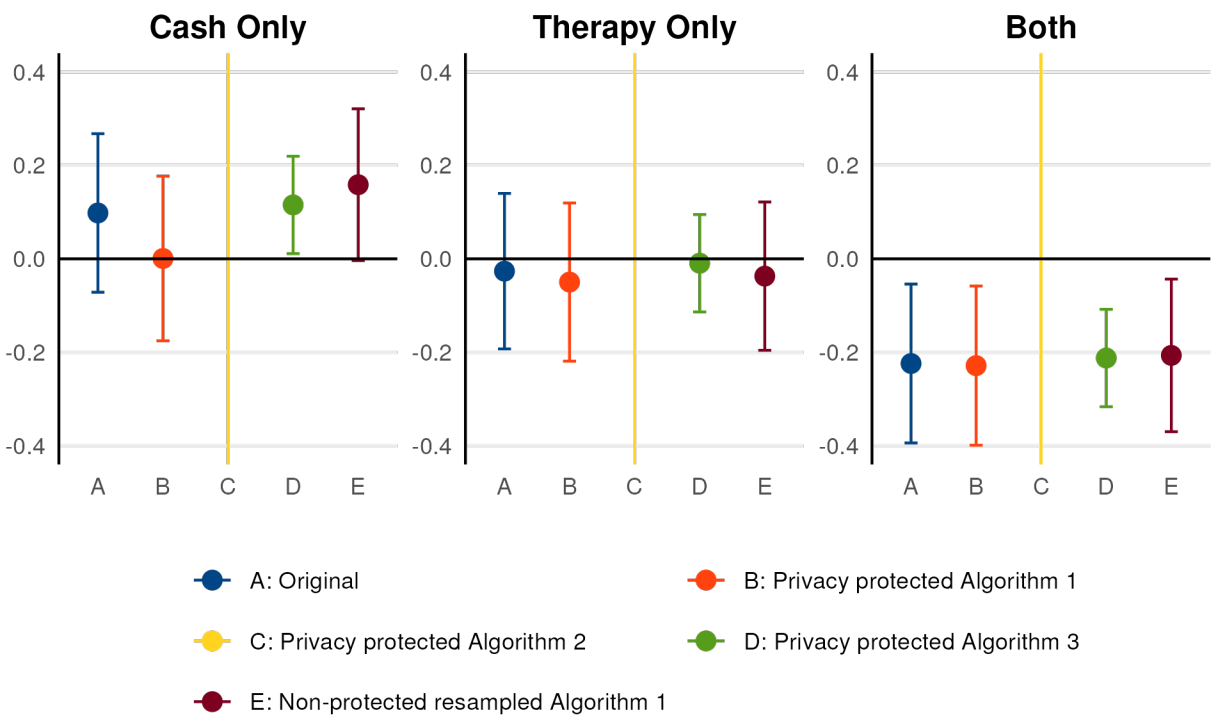

Figure 1: Outcome measures in the original study and after application of the privacy-protection measures described in the text

model-based variant of $\widetilde{Y}_3$ computed based on $\widetilde{\tau}_3$. The full paper has additional details.

3 Results

In Mukherjee et al. (2023) and Mukherjee et al. (TBD), we focus on two key tables of BJS, Table 1 — the balance table — and Table 2 — the key table in the paper, which shows the effects of the three treatment branches on outcomes, measured at two time intervals. In addition to the algorithms described earlier, we also benchmark against a variant of the base algorithm, in which data are resampled from the histogram, without any protection (i.e., $\epsilon = \infty$). In our analysis, we focus on the longer-run outcomes (Panel B of Table 2). Our analysis is not based on an exact reproduction of BJS, because a first constraint was encountered in how much memory the histogram required.[12] We limited the model to subsets of X, using between 7 and 10 columns, rather than the 13 that the authors use. Our

[12]This may be an artifact of the use of R (R Core Team, 2024), and better algorithms may exist that would alleviate this constraint.

results using the original data, however, are very close even in this reduced specification. We show that the protected data achieve similar balance as the confidential data, which is not surprising, given the design of our algorithms. Algorithms 1 and 3 obtain reasonable inference when compared to the inference based on the original data, though small differences arise. Algorithm 2 has poor inference validity, even for moderate levels of protection. Figure 1 summarizes the various outcomes for one particular coefficient across the three arms of the RCT, for various privacy-protection algorithms.[13]

In addition to the inference validity of the model when using protected data $\left[\widetilde{Y}_i, \widetilde{T}_i, \widetilde{X}_i\right]$, as depicted in Figure 1, we also assessed empirical measures of how strongly protected the data are, based on distributional comparisons of the confidential data to the protected data (MSE and KL measures). Finally, we tracked the computational burden of the algorithms. Conditional on the problem being able to fit into memory — a major constraint — the computational time was in general measured in minutes, and seemed acceptable.

4 Discussion

Much of the literature (in economics) publishes replication packages with either weakly protected (de-identified) data, or withholds the data out of privacy concerns, impeding the ability for others to investigate inference. Our goals are to maintain the ability to publish data as part of replication packages, yet provide stronger protections. A straightforward application of one of the simplest DP mechanisms (histogram count perturbation with Laplace noise) is applied to the sensitive covariate data. We show that using a hybrid approach, where the outcomes of interest (potentially not observable by the broader public) are protected with "traditional" methods, constitutes an improvement. Yet such a method does not comply with a complete differentially private protection, since some knowledge about the confidential values of the covariate data can "leak out" through the variance-covariance matrix of the released data. We therefore also explore stronger, more complex, fully differentially private

[13]Results stem from a pre-publication version of Mukherjee et al. (TBD), and the final versions may appear different in that publication.

mechanisms. While the application is more complex, they remain computationally tractable for the relatively simple dataset of our pilot study. But they do have other shortcomings: Our general purpose Algorithm 2 provides poor inference validity, whereas the more complex Algorithm 3 provides good inference validity, at the cost of not basing it on the released data, relying yet again on non-shareable confidential data. It is also non-trivial to measure the privacy loss in complex algorithms.

A researcher must therefore make a choice, possibly in conjunction with ethicists and respondents, on how much, and how strongly, to protect the data. The proposed Algorithm 1, while not completely DP, provides better protection than traditional methods, and may be acceptable in certain contexts.

Our analysis has various caveats, which we describe in detail in Mukherjee et al. (TBD). Importantly, we have relied on some key features of the typical RCT: Randomization is orthogonal to the observed covariates of participants, and thus is non-informative about those covariates. Relatively few parameters are of key interest, and estimated coefficients on other control variables are typically not published. Overall, this reduces the amount of information that needs to be released, greatly facilitating the application of strong privacy protection. Relaxing any of these features may lead to less favorable results.

Two additional issues, however, are worth highlighting here. First, our experiments suggest that in order to obtain good inference, model-based algorithms that target a particular application — the research question of the published article — must be used. This will, in turn, reduce the ability of others to use the same data in different contexts, and for others to conduct robustness tests that question the original model used. Such questions can only be answered with renewed access to the confidential data. Support for such post-publication inquiries and access is generally weak, unless researchers are working in the context of large research centers. Issues of funding for such infrastructure are a perennial problem when research centers and universities discuss data preservation policies. Our results suggest that such infrastructure would be increasingly in demand if better privacy protections are applied,

even when protected data can be published (see also Vilhuber, n.d.).

Second, we believe that the role of software packages is crucial to the wider adoption of stronger practices. The methods we developed rely on published algorithms throughout, and were simply applied to the problem at hand, yet constituted *de novo* software implementations. The development or extension of software packages that implement these and other methods, in software platforms used by the relevant social scientists (R, Stata) should be supported.[14]

The ideas and results reported here are the first step towards better understanding of feasible privacy-preservation of RCTs-based data, ensuring that privacy of data contributors to RCTs, often from low- and middle-income countries (LMIC) countries, will be more strongly protected, while maintaining the ability to draw meaningful inferences. While policy-oriented stakeholders are primarly interested in the latter, citizens that contribute their data to RCTs and companies, such as fin-tech providers, that provide key data to researchers are also heavily invested in protecting privacy. Consumer and citizen protection agencies, ethic review boards, and other regulators, should be interested in knowing of the existence of privacy-enhancing methods, possibly facilitating approval of studies in the presence of strong privacy guarantees.

[14]We note the ongoing development of non-DP methods in `synthpop`, an R package (Nowok, Raab and Dibben, 2016; Snoke et al., 2018; Raab, Nowok and Dibben, 2021).

References

Abowd, John, and Ian M. Schmutte. 2015. "Economic analysis and statistical disclosure limitation." *Brookings Papers on Economic Activity*, 221–267. https://doi.org/10.1353/eca.2016.0004.

Alabi, Daniel, Audra McMillan, Jayshree Sarathy, Adam Smith, and Salil Vadhan. 2020. "Differentially private simple linear regression." https://arxiv.org/abs/2007.05157. tex.howpublished: arXiv:2007.05157 [cs.LG] tex.optabstract: tex.optgrants: Simons Investigator Award, Cooperative Agreement CB16ADR0160001 with the Census Bureau tex.optkeywords: tex.optsource:.

Awan, Jordan, and Aleksandra Slavković. 2020. "Structure and sensitivity in differential privacy: Comparing k-norm mechanisms." *Journal of the American Statistical Association*, 1–20.

Balle, Borja, Gilles Barthe, and Marco Gaboardi. 2018. "Privacy amplification by subsampling: Tight analyses via couplings and divergences." 6280–6290. http://papers.nips.cc/paper/7865-privacy-amplification-by-subsampling-tight-analyses-via-couplings-and-divergences tex.bibsource: dblp computer science bibliography, https://dblp.org tex.biburl: https://dblp.org/rec/conf/nips/BalleBG18.bib tex.timestamp: Fri, 06 Mar 2020 17:00:31 +0100.

Barrientos, Andrés F., Alexander Bolton, Tom Balmat, Jerome P. Reiter, John M. de Figueiredo, Ashwin Machanavajjhala, Yan Chen, Charley Kneifel, and Mark DeLong. 2018. "Providing access to confidential research data through synthesis and verification: An application to data on employees of the U.S. federal government." *The Annals of Applied Statistics*, 12(2): 1124 – 1156. https://doi.org/10.1214/18-AOAS1194.

Bernanke, Ben S. 2004. "Editorial Statement." *The American Economic Review*, 94(1): 404–404. https://www.jstor.org/stable/3592790 (accessed 2020-09-01).

Blattman, Christopher, Julian C. Jamison, and Margaret Sheridan. 2017*a*. "Reducing Crime and Violence: Experimental Evidence from Cognitive Behavioral Therapy in Liberia." *American Economic Review*, 107(4): 1165–1206. https://doi.org/10.1257/aer.20150503.

Blattman, Christopher, Julian C. Jamison, and Margaret Sheridan. 2017*b*. "Replication data for: Reducing Crime and Violence: Experimental Evidence from Cognitive Behavioral Therapy in Liberia." American Economic Association [publisher] data, https://doi.org/10.3886/E113056V1.

Bowen, Claire McKay, Victoria Bryant, Leonard Burman, Surachai Khitatrakun, Robert McClelland, Philip Stallworth, Kyle Ueyama, and Aaron R Williams. 2020. "A synthetic supplemental public use file of low-income information return data: methodology, utility, and privacy implications." 257–270, Springer.

Department of Health and Human Services. 2012. "Methods for De-identification of PHI." https://www.hhs.gov/hipaa/for-professionals/privacy/special-topics/de-identification/index.html (accessed 2020-08-26).

Desfontaines, Damien, and Balázs Pejó. 2022. "SoK:Differential Privacies."

DIME. 2020. "De-identification." World Bank Dimewiki. https://dimewiki.worldbank.org/De-identification (accessed 2022-06-12).

D'Orazio, Vito, Marcus Deng, and Michael Shoemate. 2018. "TwoRavens for Event Data." 394–401. https://doi.org/10.1109/IRI.2018.00065.

Duflo, Esther. 2020. "Field Experiments and the Practice of Policy." *American Economic Review*, 110(7): 1952–73. https://doi.org/10.1257/aer.110.7.1952.

Dwork, Cynthia, and Aaron Roth. 2014. "The Algorithmic Foundations of Differential Privacy." *Found. Trends Theor. Comput. Sci.*, 9(3–4): 211–407. https://doi.org/10.1561/0400000042.

Dwork, Cynthia, Frank McSherry, Kobbi Nissim, and Adam Smith. 2006. "Calibrating Noise to Sensitivity in Private Data Analysis." *TCC'06*, 265–284. Berlin, Heidelberg:Springer-Verlag. https://doi.org/10.1007/11681878_14. tex.ids= dwork-CalibratingNoiseSensitivity2006a event-place: New York, NY.

Dwork, Cynthia, Frank McSherry, Kobbi Nissim, and Adam Smith. 2016. "Calibrating Noise to Sensitivity in Private Data Analysis." *Journal of Privacy and Confidentiality*, 7(3). https://doi.org/10.29012/jpc.v7i3.405.

Dwork, Cynthia, Weijie Su, and Li Zhang. 2021. "Differentially private false discovery rate control." *Journal of Privacy and Confidentiality*, 11(2). https://doi.org/10.29012/jpc.755.

Hundepool, Anco, Josep Domingo-Ferrer, Luisa Franconi, Sarah Giessing, Eric Schulte Nordholt, Keith Spicer, and Peter-Paul De Wolf. 2012. *Statistical disclosure control.* Vol. 2, Wiley New York.

Kancharla, Manjusha, and Hyunseung Kang. 2021. "A Robust, Differentially Private Randomized Experiment for Evaluating Online Educational Programs With Sensitive Student Data." https://doi.org/10.48550/arXiv.2112.02452.

Karwa, Vishesh, and Salil Vadhan. 2017. "Finite Sample Differentially Private Confidence Intervals."

Kazan, Zeki, Kaiyan Shi, Adam Groce, and Andrew P Bray. 2023. "The Test of Tests: A Framework for Differentially Private Hypothesis Testing." Vol. 202 of *Proceedings of Machine Learning Research*, 16131–16151. PMLR. https://proceedings.mlr.press/v202/kazan23a.html.

Kopper, Sarah, Anja Sautmann, and James Turitto. 2020. "J-PAL GUIDE TO DE-IDENTIFYING DATA." J-PAL. https://www.povertyactionlab.org/sites/default/files/research-resources/J-PAL-guide-to-deidentifying-data.pdf (accessed 2022-06-12).

Machanavajjhala, Ashwin, and Daniel Kifer. 2015. "Designing Statistical Privacy for Your Data." *Communications of the ACM*, 58(3): 58–67. https://doi.org/10.1145/2660766.

Machanavajjhala, Ashwin, Johannes Gehrke, Daniel Kifer, and Muthuramakrishnan Venkitasubramaniam. 2006. "l-Diversity: Privacy beyond k-Anonymity." 24. IEEE Computer Society. https://doi.org/10.1109/ICDE.2006.1. tex.bibsource: dblp computer science bibliography, https://dblp.org tex.biburl: https://dblp.org/rec/conf/icde/MachanavajjhalaGKV06.bib tex.timestamp: Wed, 16 Oct 2019 14:14:56 +0200.

Meager, Rachael. 2019. "Understanding the Average Impact of Microcredit Expansions: A Bayesian Hierarchical Analysis of Seven Randomized Experiments." *American Economic Journal: Applied Economics*, 11(1): 57–91. https://doi.org/10.1257/app.20170299.

Mukherjee, Soumya, Aratrika Mustafi, Aleksandra Slavković, and Lars Vilhuber. 2023. "Assessing Utility of Differential Privacy for RCTs." arXiv stat.AP 2309.14581. https://arxiv.org/abs/2309.14581.

Mukherjee, Soumya, Aratrika Mustafi, Aleksandra Slavković, and Lars Vilhuber. TBD. "Assessing Utility of Differential Privacy for RCTs." *Harvard Data Science Review*.

Nowok, Beata, Gillian Raab, and Chris Dibben. 2016. "Synthpop: Bespoke Creation of Synthetic Data in R." *Journal of Statistical Software, Articles*, 74(11): 1–26. https://doi.org/10.18637/jss.v074.i11.

Pistner, Michelle Nixon. 2020. *Privacy Preserving Methods in the Era of Big Data: New Methods and Connections.* https://etda.libraries.psu.edu/catalog/18340map5672.

Raab, Gillian M., Beata Nowok, and Chris Dibben. 2021. "Assessing, Visualizing and Improving the Utility of Synthetic Data." *arXiv:2109.12717 [stat].*

R Core Team. 2024. "R: A Language and Environment for Statistical Computing." Vienna, Austria, R Foundation for Statistical Computing. https://www.R-project.org/.

Rosenblatt, Lucas, Bernease Herman, Anastasia Holovenko, Wonkwon Lee, Joshua Loftus, Elizabeth McKinnie, Taras Rumezhak, Andrii Stadnik, Bill Howe, and Julia Stoyanovich. 2023. "Epistemic Parity: Reproducibility as an Evaluation Metric for Differential Privacy." https://doi.org/10.48550/arXiv.2208.12700.

Samarati, Pierangela, and Latanya Sweeney. 1998. "Protecting Privacy When Disclosing Information: K-Anonymity and Its Enforcement through Generalization and Suppression." Harvard University Mimeo. https://dataprivacylab.org/dataprivacy/projects/kanonymity/paper3.pdf (accessed 2023-11-08).

Seeman, Jeremy, Aleksandra Slavkovic, and Matthew Reimherr. 2020. "Private Posterior Inference Consistent with Public Information: A Case Study in Small Area Estimation from Synthetic Census Data." 323–336, Springer. https://doi.org/10.1007/978-3-030-57521-2_23.

Slavkovic, Aleksandra, and Jeremy Seeman. 2022. "Statistical Data Privacy: A Song of Privacy and Utility." https://doi.org/10.48550/ARXIV.2205.03336.

Slavkovic, Aleksandra, and Roberto Molinari. 2021. "Perturbed M-Estimation: A Further Investigation of Robust Statistics for Differential Privacy."

Snoke, Joshua, Gillian M. Raab, Beata Nowok, Chris Dibben, and Aleksandra Slavkovic. 2018. "General and Specific Utility Measures for Synthetic Data." *Journal*

of the Royal Statistical Society Series A: Statistics in Society, 181(3): 663–688. https://doi.org/10.1111/rssa.12358.

Vilhuber, Lars. 2020. "Reproducibility and Replicability in Economics." *Harvard Data Science Review*, 2(4). https://doi.org/10.1162/99608f92.4f6b9e67.

Vilhuber, Lars. 2024. "Report of the AEA Data Editor." *AEA Papers and Proceedings*, 114: 878–890. https://doi.org/10.1257/pandp.114.878.

Vilhuber, Lars. n.d.. "Using containers for analysis validation at scale." *NBER proceedings*, this issue.

Vlaeminck, Sven. 2021. "Dawning of a New Age? Economics Journals' Data Policies on the Test Bench." *LIBER Quarterly: The Journal of the Association of European Research Libraries*, 31(1): 1–29. https://doi.org/10.53377/lq.10940.

Vu, Duy, and Aleksandra Slavkovic. 2009. "Differential Privacy for Clinical Trial Data: Preliminary Evaluations." *ICDMW '09*, 138–143. Washington, DC, USA:IEEE Computer Society. https://doi.org/10.1109/ICDMW.2009.52.

Warner, Stanley L. 1965. "Randomized Response: A Survey Technique for Eliminating Evasive Answer Bias." *Journal of the American Statistical Association*, 60(309): 63–69. http://www.jstor.org/stable/2283137 (accessed 2023-11-05).

Wasserman, Larry, and Shuheng Zhou. 2010. "A Statistical Framework for Differential Privacy." *Journal of the American Statistical Association*, 105(489): 375–389. https://doi.org/10.1198/jasa.2009.tm08651. _eprint: arXiv:0811.2501v2.

5 Acknowledgement and Disclosure of funding

This work was supported through Digital Credit Observatory (CEGA, University of California, Berkeley/Bill and Melinda Gates Foundation (MP)), and in-part by NSF awards SES-1853209

and CNS-1702760 to Penn State University. We would like to thank Luqi Emanuele and Shuhang Lou for their contributions towards this work and its extensions. We declare no known conflicts of interest.