Assessing Utility of Differential Privacy for RCTs

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

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

September 12, 2022

DRAFT - Do not cite or quote

Abstract

Randomized control trials, RCT, have become a powerful tool for assessing the impact of interventions and policies in many contexts. Today, they are considered the gold-standard for inference in the biomedical fields and in many social sciences. In economics, much of the growth has been since the 1990s. Studies can involve small-scale interventions, randomized at the personal, family, or village level, but are sometimes also measured with province- or national-level outcomes. Researchers have published an increasing number of studies that rely on RCTs for at least part of the inference.

In the meantime, differential privacy (DP) has been proposed as a principled framework to privacy protection. However, there is still no sufficient guidance for social science researchers and practitioners on how to implement DP methods, and how to analyze DP-protected data. Furthermore, concerns have been expressed that DP methods may reduce utility (here: inference validity), or require much larger samples in order to achieve similar utility while providing sufficient (superior) privacy protection.

In this study, we empirically evaluate the impact of DP methods on published analyses from randomized control trials (RCTs), leveraging the availability of numerous replication packages (research compendia) in economics and policy analysis. We aim to assess, for each paper, whether an optimally chosen differentially private protection mechanism would have lead to similar inferences compared to non DP-sanitized data, or if not, what the necessary sample size augmentation may need to be; optimality criteria will be defined with respect to desired valid statistical outcomes (e.g., low type 1 error, high power, unbiasness, etc). From this analysis, we aim to distill guidance to researchers wishing to implement RCTs as to the choice between privacy protection and power in a planned study.

This proposed paper addresses the NBER conference's topic of "How does the use of privacy-protected data affect the identification, estimation, and uncertainty quantification of parameters of interest?" for field experiments and any other analysis using RCTs or surveys, in multiple fields of economics.

1 Introduction

Randomized control trials, RCT, have become a powerful tool for assessing the impact of interventions and policies in many contexts. Today, they are considered the gold-standard for inference in the biomedical fields and in many social sciences. In economics, much of the growth has been since the 1990s. Studies can involve small-scale interventions, randomized at the personal, family, or village, but are sometimes also measured with province- or national-level outcomes. Researchers have published an increasing number of studies that rely on RCTs for at least part of the inference.

In a parallel development, the improvement in transparency in the social sciences has lead to more and more of the supplementary materials for these articles to be made public as "replication packages". For instance, the American Economic Association (AEA) journals for applied economics (AEJ:Applied) and economic policy (AEJ:EP), created in 2009, have since their inception required that analysis data and code be made available. This 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. Meager (2019) re-analyzed numerous RCTs to assess the robustness of their findings using Bayesian hierarchical analysis (BHA). We will return to Meager (2019) later.

The data included in such replication packages usually allows to reproduce the results in the papers exactly, suggesting that all the analysis is conducted on these data. However, the typical guidance followed by researchers who conduct RCTs (Kopper, Sautmann and Turitto, 2020; DIME, 2020; Department of Health and Human Services, 2012) suggests primarily de-identification, the most basic anonymization, as the protection mechanism, and where further anonymization is suggested, more traditional disclosure avoidance methods (e.g., l-diversity, Machanavajjhala et al. (2006); Hundepool et al. (2012), and other aggregation-based methods are suggested). Differential privacy (DP) is sometimes referenced (Dwork et al., 2016; Wood et al., 2021), but as far as we are aware, no straightforward guidance for social science researchers and practitioners is available on how to implement DP, and how to analyze DP-protected data. This suggests that much of the current literature is based on data analysis that is public, but possibly inadequately protected. This

DRAFT - Do not cite or quote

is particularly concerning because many of these studies have data from respondents in low and middle income countries (LMIC).

One of the reasons is that there have so far not been tools available to non-specialists that would allow for easy but efficient protection using differentially private tools. Efficiency here is defined as "perturbing inference as little as possible compared to the unprotected inference." We note that inference even in the "unprotected" case is already subject to uncertainty that is often not adequately taken into account, as evidenced by Meager (2019). This is even more important in for the uncertainty and data modifications due to statistical disclosure limitation (SDL) (Abowd and Schmutte, 2015; Slavkovic and Seeman, 2022, and references therein), who demonstrate a need and (in the case of Slavkovic and Seeman, 2022) a way to make an adjustment for privacy-preservation noise in addition to other source of uncertainty.

1.1 Research questions and academic contribution

This project sets out to provide a first 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. More broadly, the project will contribute to a (short) literature on privacy-aware analysis, and privacy-aware planning for such analyses.

The project is, as far as we know, the first systematic exploratory analysis of RCTs to understand the impact of privacy-preservation and with the focus on LMIC data.

The project proposed here is innovative in two separate dimensions. First, it will assess the feasibility of stronger privacy protections for data collected in LMIC, taking into account the ability to make robust (or any) inferences. Second, it contributes to the statistical and computer sciences literature (and the fields relying on these) assessing the interaction between causal inference and privacy protection, in particular when privacy protection is conducted via DP methods.

We believe that the focus on RCTs is particularly well-suited for this endeavor, for several reasons. First, methods are, in general, quite straightforward: OLS, difference-in-difference methods, 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

DRAFT - Do not cite or quote

Slavković, 2020; Alabi et al., 2020; Slavkovic and Molinari, 2021; Barrientos et al., 2018; Bowen et al., 2020). Second, most RCTs are small-scale, using samples of the overall population, allowing us to leverage privacy-amplifying methods (Balle, Barthe and Gaboardi, 2018). 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 been also explored within the DP framework (e.g., Vu and Slavkovic, 2009; Pistner, 2020; Dwork, Su and Zhang, 2021)). Finally, 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 a convenient analogy to improve acceptance of such methods. Furthermore, a similar Bayesian framework can be used to adjust noisy inference due to privacy (e.g., Seeman, Slavkovic and Reimherr (2020).)

1.2 Research design and data.

We will re-analyze 10-12 studies that have been published, that use data from LMIC, and for which it is known that the data documentation has been reviewed and is usable without a large investment. The latter part is achieved by relying, where possible, on studies curated by our partners at *Innovations for Poverty Action* (IPA). Additional sources may be the same set of studies used by Meager (2019), which have already been re-analyzed once. A tentative initial list is provided in Table 1.

We will analyze each article, identify the analysis method used, identify the variables of interest, as well as the data generating process. Based on our knowledge of the DP literature, we will choose the most efficient data protection (for example, based on release of synthetic microdata or summary statistics) and where necessary adapted analysis method for the proposed inference. We will leverage, where possible, existing methods (using R code and packages, some our own) and emerging toolkits (e.g., openDP, Tumult lab's forthcoming system) that are well understood and accessible to other researchers. We will then re-run the analysis, and compare the protected inference to the original inference.

We note immediately a particular issue that we expect to encounter. The analysis provided

DRAFT. Do not cite or quote

1. Randomizing Religion: The Impact of Protestant Evangelism on Economic Outcomes (Dataverse entry; publication)
2. Identifying Information Asymmetries in a Consumer Credit Market (Dataverse entry; project summary)
3. Getting to the Top of Mind: How Reminders Increase Saving (Dataverse entry; working paper,publication)
4. Do Information Technologies Improve Teenagers' Sexual Education? Evidence from a Randomized Evaluation in Colombia (Dataverse entry; publication)
5. Sustainable Transformation for Youth in Liberia (Dataverse entry; project summary)
6. Profitability of Fertilizer: Experimental Evidence from Female Rice Farmers in Mali (Dataverse entry; publication)
7. The Impact of Financial Education for Youth in Ghana (Dataverse entry; working paper)
8. Partnership Schools for Liberia (Dataverse entry; project summary)

Table 1: Tentative list of RCTs with available data, LMIC

here is contingent on data that have already been collected. If, as we expect, inference based on privacy-protected data is more tenuous and noisier, one implication is that the typical power calculation should be adjusted to take privacy protection into account, presumably leading to higher required sample sizes (e.g., Vu and Slavkovic, 2009). Expressed differently, to the extent that power calculations suggested the appropriate minimal sample sizes used by the studies that we will re-analyze, these studies may be under-powered with respect to optimal privacy-protected inference. As part of our analysis, we will endeavor to recover the necessary sample size that would have been needed in order to obtain the same originally intended power.

2 Privacy analysis of chosen papers

Coming (expected February 2023).

DRAFT - Do not cite or quote

3 DP methods for chosen papers

Coming (expected April 2023)

4 Results

Expected June 2023.

5 Discussion

Coming.

The results from this project will ensure that privacy of data contributors to RCTs 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 such methods, possibly facilitating approval of studies in the presence of strong privacy guarantees.

DRAFT - Do not cite or quote

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.

Bowen, Claire McKay, Victoria Bryant, Leonard Burman, Surachai Khitatrakun, Robert McClelland, Philip Stallworth, Kyle Ueyama, and Aaron R Williams. 2020. "A synthetic

DRAFT - Do not cite or quote

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).

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

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.

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, 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.

DRAFT - Do not cite or quote

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.

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

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.

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."

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.

Wood, Alexandra, Micah Altman, Kobbi Nissim, and Salil Vadhan. 2021. "Designing Access with Differential Privacy." In *Handbook on Using Administrative Data for Research and Evidence-based Policy*. , ed. Shawn Cole, Iqbal Dhaliwal, Anja Sautmann and Lars Vilhuber. Abdul Latif Jameel Poverty Action Lab. https://doi.org/10.31485/admindatahandbook.1.0.

DRAFT - Do not cite or quote