Assessing Utility of Differential Privacy for RCTs

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

April 27, 2023

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.

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. 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. Meager (2019) re-analyzed numerous RCTs to assess the robustness of their findings using Bayesian hierarchical analysis (BHA). Roth (2022) selected event studies for which complete replication packages were available, to re-analyze them in light of pre-treatment time treands. These kinds of studies are possible because of the increased availability of complete replication materials.[1]

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 (Department of Health and Human Services, 2012; Kopper, Sautmann and Turitto, 2020; DIME, 2020) 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

[1] It should be noted that Roth (2022) still had to exclude nearly four times as many papers as they included because data were not readily available.

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 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 for the uncertainty and data modifications that are generated through statistical disclosure limitation (SDL). Abowd and Schmutte (2015); Slavkovic and Seeman (2022) demonstrate the need to account for the privacy-preserving noise in analyses. Slavkovic and Seeman (2022) propose 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 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. More broadly, the project will contribute to a 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 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 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 also been 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 analyze several previously published studies with complete available data. These may come from the aforementioned journals with a robust data and code availability policies, or from studies that have been separately verified by institutions active in the domain of LMIC RCTs, such as *Innovations for Poverty Action* (IPA), J-PAL, or 3ie, all of which have or have had active reproducible checks.

We analyze each article, identify the analysis method used, identify the variables of interest, as well as the data generating process. Based on a review of the DP literature, we 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,) and emerging toolkits (e.g., openDP, Tumult Lab's 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 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 Problem setup

The standard problem setup is described here. The experimenter is interested in determining whether a particular treatment has any effect on a response variable when the treatment is applied to an entity, individual, or treatment unit. The typical manner in the experiment is performed to answer the experimenter's query is to randomly assign the treatment to the treatment units according to some chosen experimental design, apply the treatment to the units and measure/record the response variable after the treatment is applied to the units. The statistical analysis is based on regressing the response variable on the levels of the treatment applied and then inferring about the effect of the treatment on the response. In order to better understand how this effect may or may not vary based on inherent characteristics of the treatment units, the experimenter also typically records or measures additional variables (covariates, control variables) and accounts for these variables in the regression model in order to improve the statistical utility of the estimate of the treatment effect and the power of testing procedures concerning the treatment effect. However, these covariates pose a privacy concern for the individuals or treatment units participating in the study. An (privacy) attacker, if provided with the database containing the covariate information, for instance from a replication package, may link some records in the given database with an external database, and thus gain knowledge of characteristics of one or more treatment units, along with the level of treatment

received by the concerned treatment units, which the attacker did not possess before the database was provided to her. This constitutes a privacy violation of the treatment units.

There are two types of competing factors at play. The experimenter is responsible for providing privacy protection to the participating entities in a randomized controlled trial, ideally by using methodology that satisfies formal privacy guarantees. But the experimenter also wishes to maximize the utility of the randomized controlled trial with respect to the scientific knowledge it generates. There are two ways an experimenter aims to make an RCT scientifically useful. Firstly, the experimenter performs a statistical analysis of the data collected, reports the results of the analysis in a summarized form and most importantly, infers the effect of the treatment on the response variable (typically by means of a point/interval estimate or a hypothesis test). Secondly, the experimenter makes the data required for the statistical analysis available to the public for the purpose of reproducibility, transparency and promotion of further research. The publication of the analysis results along with the analysis data in the public domain has the potential to violate privacy. On the other hand, perturbing either the summary statistics or the analysis data before publication in order to provide privacy protection reduces their statistical utility and the reproducibility of the research performed by the experimenter.

3 Data structure and Goal

We consider the scenario where the experimenter/analyst is interested in determining the main effect of one or more treatment variables using a regression model with fixed effects and each treatment variables have a finite number of treatment levels associated with it. In addition, there may be additional variables which are used for stratification or blocking. If there is no stratification involved, it is assumed that the experimental design uses a simple randomization scheme where the treatment units are assigned to the treatment level combinations using simple random sampling with replacement (we can perform without replacement sampling as well, but for simplicity we focus on the sampling with replacement scenario). If there are one or more blocking variables involved, it is assumed that the experimental design uses a randomization scheme where the treatment units

within a block (corresponding to a unique combination of the blocking variables) are assigned to the treatment level combinations using simple random sampling with replacement. This experimental design is referred to as a randomized block design.

The covariates incorporated in the regression model in addition to the treatment and blocking variables can be either discrete or continuous. If a covariate is discrete, it is assumed that that the number of distinct values of each discrete covariate is finite and these distinct values are either exactly known or belong to a known finite set.

The analysis data is assumed to be available in the form of a dataframe with n rows and $p+t+b+1$ columns, where n is the total number of treatment units, t is the number of dummy variables required for representing all possible treatment combination assignments using dummy coding and b is the number of blocking variables. Note that we consider the treatment units in the control group as treatment units which are assigned to a particular treatment level combination. The i-th row corresponds to the i-th treatment unit, $i = 1, \ldots, n$. The first column contains the values of the response variable y. The next t columns represent the dummy variables which indicate the treatment level combinations assigned to the treatment units. The next b columns contain the block assignments based on the b blocking variables (note that we are not assuming these block assignments to be in dummy coding form). The remaining p columns contain the data corresponding to the covariates to be included in the regression model. We refer to the last p columns as the covariate dataframe, which is the source our potential privacy concern.

The experimenter's goal is to accurately infer the main effect(s)of the treatment variable(s) using a regression model with fixed effects. In addition release the dataframe containing the analysis data in the public domain.

Our goal is to reduce the privacy violation risk using a differentially private data release mechanism that allow the sanitized dataframe to be released without compromising the quality of inference about the parameter(s) of interest. More specifically, the data release mechanism must ensure differential privacy at the level of individual treatment units. Further, it must ensure that the statistical inference (for example point estimation, interval estimation or hypothesis testing) regarding the

main effect(s) of the treatment variable(s) performed using the sanitized dataset is very "similar", in a way that we will define shortly, to the inference performed using the original dataframe which contains the private information is not sanitized before release.

4 Synthetic data generation approach based on perturbed histogram

Assuming that a linear regression model is suitable, the regression model of interest in the absence of blocking variables is given by

$$y_i = \alpha + \sum_{k=1}^{b} \tau_k T_{k,i} + \sum_{l=1}^{p} \gamma_l X_{l,i} + \epsilon_i, \quad i = 1, \ldots, n \tag{1}$$

where T_k represent the dummy variables for the treatment level combinations and X_l represent the covariates/control variables associated with the n treatment units and $\epsilon_i \overset{i.i.d}{\sim} N(0, \sigma^2)$.

When stratification is used with a total of m block combinations and n_j treatment units are assigned to j-th block combination, the corresponding regression model is given by

$$\begin{gathered} y_{ij} = \alpha + \sum_{k=1}^{b} \tau_k T_{k,i} + \sum_{l=1}^{p} \gamma_l X_{l,ij} + \epsilon_{ij} \\ i = 1, \ldots, n_j, j =, 1, \ldots, m, \sum_{j}^{m} n_j = n \end{gathered} \tag{2}$$

In both the above models, the parameter(s) of interest to the experimenter are the fixed effects τ_k, $k = 1, \ldots, b$. From the point of view of the experimenter, statistical utility will be preserved if the inference concerning the fixed effects τ_k is affected as little as possible by the data release mechanism used to sanitize the analysis data in order to protect privacy.

We will adopt a synthetic data generation approach that will attempt to preserve the inference concerning the fixed effects τ_k while ensuring that the data release mechanism releases a synthetic dataframe with N observations and satisfies ϵ-differential privacy (DP).[2]

The basic idea is to extract the covariate information from the analysis data contained in the

[2]We note that it is not strictly necessary to output the exact same N observations as in the private data frame, but this seems to be the convention.

last p columns of the private dataframe and construct a generative model for the covariate data using a p-multidimensional/multivariate histogram. The histogram counts are then sanitized using the multidimensional Laplace mechanism which adds Laplace nose with mean 0 and variance $2/\epsilon$ to each count. Then, we perform treatment level (and blocking, if present) assignments using the experimental design on N synthetic treatment units. Next, the point estimates of the regression coefficients τ_k and γ_l along with the point estimate of the residual variance σ^2, which are denoted by $\hat{\tau_k}$,$\hat{\gamma_l}$ and $\hat{\sigma}^2$ are computed using the private dataframe. Once the covariate data and the treatment (and block) assignments are synthetically generated, we use the regression model (1) (accordingly (2), if blocking is present) as a generative model for the response variable y.

Note that there is essentially no restriction in extending this approach to other regression models (such as logistic regression) which might be more suitable than linear regression in some scenarios.

4.1 Algorithm

We describe the algorithm for the case where there are no blocking variables. The only change for the case where there are blocking variables is in the experimental design used to assign treatment levels and block combinations to the N synthetic treatment units, which is straightforward. The basic algorithm is based on the following steps:

1. Construct a multivariate histogram for the p-dimensional covariate data. Number of bins along each of the dimensions corresponding to the continuous variables is taken to be of the order $n^{2/3}$ and number of bins along the dimensions corresponding to the discrete variables is equal to the known number of distinct values of the variable. Let m be the number of bins required to construct the histogram. Let C_i be the count/frequency of the observations in the covariate dataframe corresponding to the i-th bin, $i = 1, \ldots, m$. Let C be the vector of counts given by $C = (C_1, \ldots, C_m)$.

2. Draw m i.i.d observations $Z_1, \ldots, Z_m$ from a Laplace distribution with location parameter/mean 0 and variance $8/\epsilon^2$ (equivalently scale parameter $2/\epsilon$). Compute the sanitized vector of counts $D = (D_1, \ldots, D_m)$ where $D_i = C_i + Z_i$, $i = 1, \ldots, m$. Since some of the sani-

tized counts could be negative valued, we transform the negative counts to 0 and renormalize the counts to obtain a vector of sanitized relative frequencies as $\widetilde{D} = (\widetilde{D}_1, \ldots, \widetilde{D}_m)$ where $\widetilde{D}_i = \frac{D_i \mathbf{I}_{D_i>0}}{\sum_{i=1}^{m} D_i \mathbf{I}_{D_i>0}}$, $i = 1, \ldots, m$.

3. Draw N i.i.d p-dimensional vectors $\widetilde{X}_1, \ldots, \widetilde{X}_N$ using simple random sampling with replacement from the m bins of the constructed histogram in Step 1 using the sanitized relative frequency vector $\widetilde{D}$ as the corresponding probabilities of each of the m bins. The sanitized covariate dataframe is denoted by $\widetilde{X}^{N\times p} = \left[\widetilde{X}_1^T \ldots \widetilde{X}_N^T\right]^T$.

4. Construct the t dummy variables corresponding to the treatment assignments using the experimental design and denote it by $\widetilde{T}^{N\times t} = \left[\widetilde{T}_1 \ldots \widetilde{T}_t\right]$. The synthetic dataframe corresponding to the treatment level assignment dummy variables and the covariates is denoted as $\widetilde{M} = [\widetilde{T}, \widetilde{X}]$.

5. Compute $\hat{\tau}_k$,$\hat{\gamma}_l$ and $\hat{\sigma}^2$ based on linear regression analysis using the original dataframe (without any privatization). (We can generalize this to any regression model).

6. Construct $\widetilde{Y} = (\widetilde{Y}_1, \ldots, \widetilde{Y}_N)$ using the privately computed $\hat{\tau}_k$,$\hat{\gamma}_l$ and $\hat{\sigma}^2$ using

$$\widetilde{Y}_i = \widetilde{M}\hat{\beta} + Z_i$$

where $Z_i \overset{i.i.d}{\sim} N(0, \hat{\sigma}^2)$, $i = 1, \ldots, N$. (We can generalize this to any prediction model based on estimated regression model).

7. Release $\widetilde{D} = [\widetilde{Y}, \widetilde{M}] = [\widetilde{Y}, \widetilde{T}, \widetilde{X}]$.

The proof of differential privacy guarantee is based on Proposition 1 in Dwork et al. (2006) along with the post-processing property of pure differential privacy, while the statistical optimality is based on Theorem 4.4 of Wasserman and Zhou (2008).

5 Numerical Experiments

In this section, we evaluate the performance of our proposed algorithm using two simulation studies. We then apply the algorithm to real-world applications in the next section. The aim of applying

the proposed privatized algorithm to a given dataset is twofold. The first aim is to ensure that any statistical inference regarding the treatment effects under study deviate as little as possible from the inference regarding the treatment effects one would obtain based on the unsanitized original dataset, while the second aim is to ensure that any statistical inference regarding the sensitive information (the covariate data) in the dataset based on the sanitized dataset is sufficiently different from the inference regarding the covariate data one would obtain based on the unsanitized dataset. The former and the latter aims will be referred to as Aim 1 and Aim 2, respectively.

Given an unsanitized dataset $D = [Y, M]$. and a sanitized version of the same dataset (synthetic dataset) $\widetilde{D} = [\widetilde{Y}, \widetilde{M}]$ obtained using our proposed algorithm for a given privacy budget ϵ, we compute the following four metrics of comparison to verify whether Aim 1 is achieved :

1. **Metric 1 - C.I. overlap indicator:** This binary (0 or 1) metric computes whether there is any overlap between the 95% confidence intervals (C.I.) for the regression coefficients (individual C.I.'s for each regression coefficient) computed based on the unsanitized dataset and the sanitized dataset.

2. **Metric 2 - Estimate coverage by sanitized C.I. indicator:** This binary (0 or 1) metric computes whether the point estimates for the regression coefficients computed based on the unsanitized dataset fall within the confidence intervals for the regression coefficients computed based on the sanitized dataset. A value of 1 indicates that the deviation of the inference regarding the regression coefficients based on the unsanitized dataset from the same inference based on the sanitized dataset is likely to be small.

3. **Metric 3 - C.I. overlap measure:** This metric computes a measure of the overlap between the 95% confidence intervals (C.I.) for the regression coefficients (individual C.I.'s for each regression coefficient) computed based on the unsanitized dataset and the sanitized dataset (Karr et al., 2006). Specifically, having chosen a particular regression coefficient β, if (L, U) is the C.I. for β computed based on the unsanitized dataset and $(\widetilde{L}, \widetilde{U})$ is the C.I. for $\widetilde{\beta}$ computed based on the sanitized dataset. Let $L^{over} = \max(L, \widetilde{L})$ and $U^{over} = \min(U, \widetilde{U})$. Then the

average overlap in confidence intervals $\widetilde{O}$ is

$$\widetilde{O} = \frac{1}{2}\left[\frac{U^{over} - L^{over}}{U - L} + \frac{U^{over} - L^{over}}{\widetilde{U} - \widetilde{L}}\right].$$

This metric is a continuous measurement version of Metric 1. The average overlap $\widetilde{O}$ can vary between 0 and 1, with higher values near 1 indicating that there is a large degree of overlap. Thus, higher values (near 1) indicate that the deviation of the inference regarding the regression coefficients based on the unsanitized dataset from the same inference based on the sanitized dataset is small.

4. **Metric 4 - Empirical Squared Error in Estimate:** This metric computes $(\beta - \widetilde{\beta})^2$, the square of the difference between the unsanitized and sanitized point estimates of the regression coefficients. Smaller values (near 0) indicate that the deviation of the inference regarding the regression coefficients based on the unsanitized dataset from the same inference based on the sanitized dataset is small.

In order to verify whether Aim 2 is satisfied, we choose a statistic that depends only on the sensitive data (the covariate data). We then compute the value of the statistic based on the unsanitized dataset and the sanitized dataset. The metric of comparison, which we refer to as **Metric 5 (Empirical Squared Error in Sensitive Statistic)**, if the squared difference between the two values of the statistic computed.

In order to obtain a measure of the performance of the proposed algorithm that takes into account the randomness in the algorithm, we compute these 5 metrics for multiple independently generated synthetic datasets and report their arithmetic mean (average). Thus, metrics 1 and 2 will be reported as proportions, and metrics 3,4 and 5 will be reported as average, when averaged over multiple synthetic datasets. Even though we simulate multiple synthetic datasets corresponding to each unsanitized dataset, we report the average metrics to obtain an indication of the performance of a single application of our proposed algorithm to obtain a single synthetic dataset, which is what we expect to be done in practice.

There are two separate sources of noise addition to the original private dataset. The first source is the statistical noise introduced due to the uncertainty involved in estimating the distribution of the covariate data and the sampling of the synthetic dataset using the estimated model. The second source is due to differential privacy (addition of Laplace noise). To assess the individual effect of noise from the second source separated from the first source, we perform the same synthetic data generation process, but without the addition of DP noise to the histogram counts (Step 2), creating a synthetic unsanitized dataset $D^* = [Y^*, M^*]$. We then again calculate the above four metrics, using D^* instead of $\widetilde{D}$ as the comparison. We finally compare the metric values with the ones we obtain from the proposed differentially private synthetic data generation process. Thus, we obtain an idea of the individual effect of the differential privacy constraint on our data generation process. Additionally, if the comparison metric values for the DP and non-DP procedures do not differ very much, we would prefer to implement the DP procedure in practice. This is because the non-DP method is vulnerable to reconstruction attacks and other forms of loss in privacy, due to its use of unsanitized histogram counts, and the loss in utility due to the additional noise injected due to the sanitization is not significant (at least on an empirical basis).

5.1 Simulation Study 1

For our first simulation study we consider a dataframe with $n = 100$ observations, 1 treatment variable with two treatment levels, "0" and "1" denoting whether or not the treatment was applied to the corresponding treatment unit and $p = 1$ continuous covariate, where we considered two different distributions for the continuous covariate: Uniform(-5,5) and Beta(1,2). The treatment variable is generated from a binomial distribution with equal probabilities for the two treatment levels. All variables are generated independent of each other. We choose the true regression coefficient as $\alpha = 0.05$ (Intercept term),$\tau_1 = 1, \gamma_1 = 0.2$ and the true residual variance to be 0.5. We denote the response variable as y, the treatment variable as x_1 and the single covariate as x_2.

We consider 3 different choices of the privacy budget ϵ as 0.1, 0.5 and 1. For a given privacy budget, we simulate 100 different datasets (response variable, treatment variable and covariate combined). For each of these 100 datasets, we independently generate 20 synthetic datasets using

our proposed algorithm and we use the chosen privacy budget for each of these applications of the algorithm. For each simulated unsanitized dataset, we estimate the model parameters using observe the deviation in inference between the unsanitizied dataset and the synthetically generated datasets in terms of the five metrics.

We run 100 simulations for 3 different choices of the privacy budget ϵ and observe the change in inference between the original simulated dataframe (which needs to be protected) and the synthetically generated dataframe. We consider the OLS point estimates and confidence intervals for the regression coefficients when computing the Metrics 1,2,3 and 4 to measure the degree of preservation of utility of the inference, not only for the treatment effects but also for the other regression coefficients. Further, to compute Metric 5, we choose the variance of the covariate x_2 as the sensitive statistic that depends on the sensitive covariate data.

5.1.1 Results of Simulation Study 1:

We report the results for Simulation study 1 using both the uniform covariate and the beta covariates. We first discuss the results when the uniform covariate is used.

In Tables 1, 2, 3, we compute the metric values for three different choices of the privacy budget $\epsilon = 0.1, 0.5$ and 1. We observe that Metric 1 always have value 1 indicating that in all these cases there is always an overlap between the original and synthetic data. The values under Metric 2 indicate that in all the cases, for all the regression coefficients, around $94 - 95\%$ of the time the value of the point estimate of the original/unsanitized dataset lies within the confidence intervals for the regression coefficients computed based on the synthetic dataset. From the values under Metric 3 we can conclude that the measure of overlap between the confidence intervals of the original and synthetic dataset is around $78 - 79\%$. From the values under Metric 4, we observe that the square of the differences between the point estimates of the regression coefficients based on the unsanitized dataset and the sanitized dataset are quite small. We observe that, irrespective of the privacy budget ϵ, the effect of the privatization on the utility of the estimates of the regression parameters is quite small. Thus, we can conclude on the basis of these results that the utility of the inference regarding the treatment effects as well as the remaining regression coefficients are is preserved to a large extent

even under privatization using our proposed algorithm.

Variable names	Metric 1	Metric 2	Metric 3	Metric 4
(Intercept)	1.00000	0.95000	0.79427	0.01127
x_1	1.00000	0.94650	0.79718	0.02099
x_2	1.00000	0.95350	0.78564	0.00076

Table 1: Effect on inference regarding regression coefficients measured using Metrics 1-4 for Simulation Study 1 with uniform covariate, averaged over 100 simulations of the sensitive dataframe, using 20 independently generated synthetic dataframes (with privacy budget $\epsilon = 0.1$) for each sensitive dataframe.

Variable names	Metric 1	Metric 2	Metric 3	Metric 4
(Intercept)	1.00000	0.95450	0.79703	0.01054
x_1	1.00000	0.94600	0.79684	0.02094
x_2	1.00000	0.94750	0.79166	0.00069

Table 2: Effect on inference regarding regression coefficients measured using Metrics 1-4 for Simulation Study 1 with uniform covariate, averaged over 100 simulations of the sensitive dataframe, using 20 independently generated synthetic dataframes (with privacy budget $\epsilon = 0.5$) for each sensitive dataframe.

Variable names	Metric 1	Metric 2	Metric 3	Metric 4
(Intercept)	1.00000	0.95000	0.79809	0.01046
x_1	1.00000	0.94900	0.79737	0.02094
x_2	1.00000	0.95700	0.79582	0.00065

Table 3: Effect on inference regarding regression coefficients measured using Metrics 1-4 for Simulation Study 1 with uniform covariate, averaged over 100 simulations of the sensitive dataframe, using 20 independently generated synthetic dataframes (with privacy budget $\epsilon = 1$) for each sensitive dataframe.

Variable names	Metric 1	Metric 2	Metric 3	Metric 4
(Intercept)	1.00000	0.95400	0.80176	0.00987
x_1	1.00000	0.94900	0.79460	0.02119
x_2	1.00000	0.95500	0.79557	0.00064

Table 4: Effect on inference regarding regression coefficients measured using Metrics 1-4 for Simulation Study 1 with uniform covariate, averaged over 100 simulations of the sensitive dataframe, using 20 independently generated Non-DP synthetic dataframes for each sensitive dataframe.

On the other hand, we expect that as the privacy budget ϵ decreases, we expect larger degrees of distortion of the covariate data in the synthetic data generation process. Thus, we should expect

larger differences (as ϵ decreases) between the values of sensitive statistics (which depend on the sensitive covariate data and which we aim to provide privacy protection) when computed using the unsanitized dataset and the sanitized dataset. From Table 5, we observe that the squared differences between the sensitive statistic (which we chose to be the variance of x_2) based on the unsanitized dataset and the sanitized dataset are increases as the privacy budget ϵ decreases. Other choices of the sensitive statistic also yield similar results. Thus, we conclude that both Aim 1 and Aim 2 are satisfied to a large extent, based on this simulation study using uniform covariates.

Table 4 provides us with the four metric values based on the synthetic data generated from the non-DP method which does not sanitize the histogram counts. Comparing it with tables 1, 2, 3 which are computed using the differentially private data generation algorithm, we see that the values almost similar. Thus, we can conclude empirically that our proposed method helps us provide DP guarantees without much extra cost. In Table 4, we compute the metrics 1-4 based on the data generated from the non-DP method. We see that the values are similar to the values computed based on the differentially private data generation procedure, thus empirically proving our conclusion that adding differential privacy guarantees is not coming at much extra cost. Further, in Table 5 we compute Metric 5 (MSE) for the sensitive statistic (Variance of Age) based on both DP and non-DP synthetic data generation procedures. The larger value of Metric 5 using DP synthetic data generation in comparison to the smaller value using non-DP synthetic data generation is indicative of the additional distortion introduced by the privatization.

Privacy Budget	ϵ **=0.1**	ϵ **=0.5**	ϵ **=1**	**Non-DP Synthesis**
MSE of Variance of x_2	6.888821	2.388792	1.273375	0.594822

Table 5: Effect on value of sensitive statistic (based on covariate data) measured using Metric 5 (MSE) for Simulation Study 1 using uniform covariate. Results are reported for DP synthesis with varying privacy budget ϵ and non-DP synthesis, each type of synthesis being averaged over 100 simulations of the sensitive dataframe, using 20 independently generated synthetic dataframes for each sensitive dataframe.

The results for the simulation study using the beta covariate are qualitatively very similar to the results we obtain using the uniform covariate. We report the results for the simulation study using the beta covariate in the Appendix (A).

5.2 Simulation Study 2

In our second simulation study, we consider a more generalized setup. We consider a dataframe with $n = 100$ observations, 1 treatment variable with two treatment levels, "0" and "1" denoting whether or not the treatment was applied to the corresponding treatment unit and $p = 3$ covariates. The treatment variable is generated from a binomial distribution with equal probabilities for the two treatment levels. Among the three covariates, one is a discrete variable with three distinct levels while the remaining two are continuous variables. The categorical variable (Covariate 2) is generated from a trinomial distribution with probability parameter 0.2, 0.3 and 0.5. The continuous covariates are generated from Uniform(-5,5) and Beta(1,2), denoted as Covariates 1 and 3 respectively. We choose the true regression coefficient as $\alpha = 0.05, \tau_1 = 1$ (corresponds to treatment effect), $\gamma_1 = 0.2, \gamma_2 = 0.4, \gamma_3 = 0.3$ (γ_i corresponds to Covariate i, $i = 1, 2, 3$) and the true residual variance to be 0.5. We denote the response variable as y, the treatment variable as x_1 and 3 covariates as x_2, x_3 and x_4.

As in Simulation Study 1, we consider the same three choices of the privacy budget ϵ as 0.1, 0.5 and 1, and proceed as before. We consider the OLS point estimates and confidence intervals for the regression coefficients when computing the Metrics 1, 2, 3, and 4 to measure the degree of preservation of utility of the inference, not only for the treatment effects but also for the other regression coefficients. Further, to compute Metric 5, we choose the variance of the covariate x_2 as the sensitive statistic that depends on the sensitive covariate data.

5.2.1 Results of Simulation Study 2:

We compute and report the same quantities as in Simulation Study 1 using the Tables 6, 7, 8 and 10. As expected, for all the different epsilon values, the differences between the estimates of the original data and the synthetic data is quite small, implying that privatization hasn't significantly affected the inference about the regression parameters. Table 9 provides us with the four metric values based on the synthetic data generated from the non-DP method. Comparing it with tables 6, 7, 8 which are computed using the differentially private data generation algorithm, we again see that the values

almost similar. Thus, in this study as well, we can conclude empirically that our proposed method helps us provide DP guarantees without much extra cost.

Variable names	Metric 1	Metric 2	Metric 3	Metric 4
(Intercept)	1.00000	0.94850	0.79996	0.15722
x_1	1.00000	0.94150	0.78922	0.02228
x_2	1.00000	0.95200	0.79785	0.00064
x3	1.00000	0.94900	0.80107	0.00788
x4	1.00000	0.95400	0.80560	0.07509

Table 6: Effect on inference regarding regression coefficients measured using Metrics 1-4 for Simulation Study 2 with 3 covariates, averaged over 100 simulations of the sensitive dataframe, using 20 independently generated synthetic dataframes (with privacy budget $\epsilon = 0.1$) for each sensitive dataframe.

Variable names	Metric 1	Metric 2	Metric 3	Metric 4
(Intercept)	1.00000	0.94850	0.80184	0.15304
x_1	1.00000	0.94150	0.78773	0.02228
x_2	1.00000	0.95200	0.79711	0.00063
x3	1.00000	0.94750	0.80363	0.00769
x4	1.00000	0.95250	0.80463	0.07551

Table 7: Effect on inference regarding regression coefficients measured using Metrics 1-4 for Simulation Study 2 with 3 covariates, averaged over 100 simulations of the sensitive dataframe, using 20 independently generated synthetic dataframes (with privacy budget $\epsilon = 0.5$) for each sensitive dataframe.

Variable names	Metric 1	Metric 2	Metric 3	Metric 4
(Intercept)	1.00000	0.94950	0.80237	0.15409
x_1	1.00000	0.94250	0.78814	0.02238
x_2	1.00000	0.94850	0.79161	0.00066
x3	1.00000	0.93950	0.80188	0.00790
x4	1.00000	0.96100	0.80529	0.07412

Table 8: Effect on inference regarding regression coefficients measured using Metrics 1-4 for Simulation Study 2 with 3 covariates, averaged over 100 simulations of the sensitive dataframe, using 20 independently generated synthetic dataframes (with privacy budget $\epsilon = 1$) for each sensitive dataframe.

Variable names	Metric 1	Metric 2	Metric 3	Metric 4
(Intercept)	1.00000	0.95100	0.79495	0.18993
x_1	1.00000	0.95400	0.79621	0.02058
x_2	1.00000	0.95500	0.80086	0.00062
x3	1.00000	0.95500	0.79791	0.00881
x4	1.00000	0.95250	0.78676	0.09785

Table 9: Effect on inference regarding regression coefficients measured using Metrics 1-4 for Simulation Study 2 with 3 covariates, averaged over 100 simulations of the sensitive dataframe, using 20 independently generated Non-DP synthetic dataframes for each sensitive dataframe.

Privacy Budget	ϵ =0.1	ϵ =0.5	ϵ =1	Non-DP Synthesis
MSE of Variance of x_2	1.189871	1.226455	1.173358	0.597624

Table 10: Effect on value of sensitive statistic (based on covariate data) measured using Metric 5 (MSE) for Simulation Study 2 using 3 covariates. Results are reported for DP synthesis with varying privacy budget ϵ and non-DP synthesis, each type of synthesis being averaged over 100 simulations of the sensitive dataframe, using 20 independently generated synthetic dataframes for each sensitive dataframe.

6 Application to "Reducing Crime and Violence: Experimental Evidence from Cognitive Behavioral Therapy in Liberia" (Blattman, Jamison and Sheridan, 2017)

6.1 Setup

In order to practically apply our proposed method to a real-world randomized control trial, we chose the analyses as published in Blattman, Jamison and Sheridan (2017). The associated replication files, including the de-identified data, are available in Blattman, Jamison and Sheridan (n.d.). For this application, picked a simplified version of the results reported in Table 2 Panel B of Blattman, Jamison and Sheridan (2017). The analysis data is obtained from the file named `STYL_Final.dta` as provided in the replication package (Blattman, Jamison and Sheridan, n.d.). Specifically, we look at the long term effect of therapy and cash grant (12-13 months after the program) on a summary index of antisocial behaviours (referred to as `fam_asb_lt`) exhibited by a sample of 999 high-risk youths in Monrovia, Liberia. A 2×2 factorial design is used with two blocking/stratification variables based on the groups the youths were together in when they were randomly assigned the treatments, once

at the time of being assigned to therapy (there were 55 such groups), and once at the time of being assigned to receive cash grant of 200 USD (there were 20 such groups). The treatment assignments are encoded using 3 binary treatment variables `cashassonly` (indicating whether only cash grant is received), `tpassonly` (indicating whether only therapy is received) and `tpcashass` (indicating whether both therapy and cash grant is received). The therapy assignment based blocking variable is `tp_strata_alt` while the cash grant assignment based blocking variable is `cg_strata`. In addition to the treatment variables and the blocking variables, we chose to include 7 covariates in our regression: `age_b`, `asbhostil_b`, `drugssellever_b`, `drinkboozeself_b`, `druggrassself_b`, `harddrugsever_b`, `steals_b`. The first 2 covariates are the age and antisocial behaviour index (Barret ASB and Hostility z-score) for the individuals participating in the study. These are continuous variables. The remaining covariates record the antisocial behaviour of the youths in terms of ever having sold drugs, whether they drink alcohol, whether they smoke grass/opium, whether they have ever consumed hard drugs and whether they have exhibited stealing behaviour in the 2 weeks prior to their interview, respectively. The values of these covariates are recorded to be 1 if the answer is affirmative, otherwise 0.

For convenience, we will rename the variables as shown in Table 11.

Original variable names	**Renamed variables**
fam_asb_lt	ASB family index
cashassonly	Cash Only
tpassonly	Therapy Only
tpcashass	Both
tp_strata_alt	Therapy Block
cg_strata	Cash Block
age_b	Age
asbhostil_b	Barret ASB index
drugssellever_b	Drugs Sell indicator
drinkboozeself_b	Alcohol self indicator
druggrassself_b	Grass/Opium self indicator
harddrugsever_b	Hard Drugs indicator
steals_b	Steal self indicator

Table 11: Renaming variables in Liberia study

6.2 Results of real world application

In Figure 1 and Table 12, we show the effect of using the differentially private synthetic data generation procedure (with privacy budget $\epsilon = 1$) and the non-DP synthetic data generation procedure (without sanitizing the histogram counts) on the statistical inference regarding the treatment effects in the study of interest, which are the treatment effects corresponding to the Cash Grant Only treatment, Therapy Only treatment and Both Cash and Therapy treatment. Specifically, we compute the treatment effect estimates (represented by dots in Figure 1 and reported in Table 12), the standard error of the treatment effect estimates (reported in able 12), the 95% confidence interval for the treatment effects (represented by intervals/errorbars in Figure 1) and the p-value for the individual tests of significance of the treatment coefficients (value reported in the Figure 1). We first compute these regression statistics on the true/original dataset. Then we privatize the dataset by generating a synthetic dataset with privacy budget $\epsilon = 1$ once, and then compute the regression statistics based on the synthetic/privatized dataset. Note that, since the privatization occurs via a randomized algorithm, we will obtain slightly different results when we apply the privatization repeatedly. Here, we report the result of a single instance of privatization using our proposed procedure. In addition, we also generate a non-DP synthetic dataset which differs from the DP algorithm only in the fact that there is no sanitization of the histogram counts. We observe that the inference based on the original dataset and the synthetic dataset is almost the same with respect to the treatment effects.

Next, to see the average performance of the DP as well as non-DP algorithm, across multiple data generations, for three different choices of the privacy budget $\epsilon = 0.1, 0.5$ and 1, we study the following four tables. In Tables 13, 14, 15 and 17, we report the same metrics as before. The variance of Age is taken as the sensitive statistic (which depends on the sensitive covariate data) for evaluating Metric 5.

As in the simulation studies, the differences between the estimates of the original data and the synthetic data is quite small, implying that privatization has not significantly affected the inference about the regression parameters. In Table 16, we compute the metrics 1-4 based on the data generated from the non-DP method. We see that the values are similar to the values computed based on the

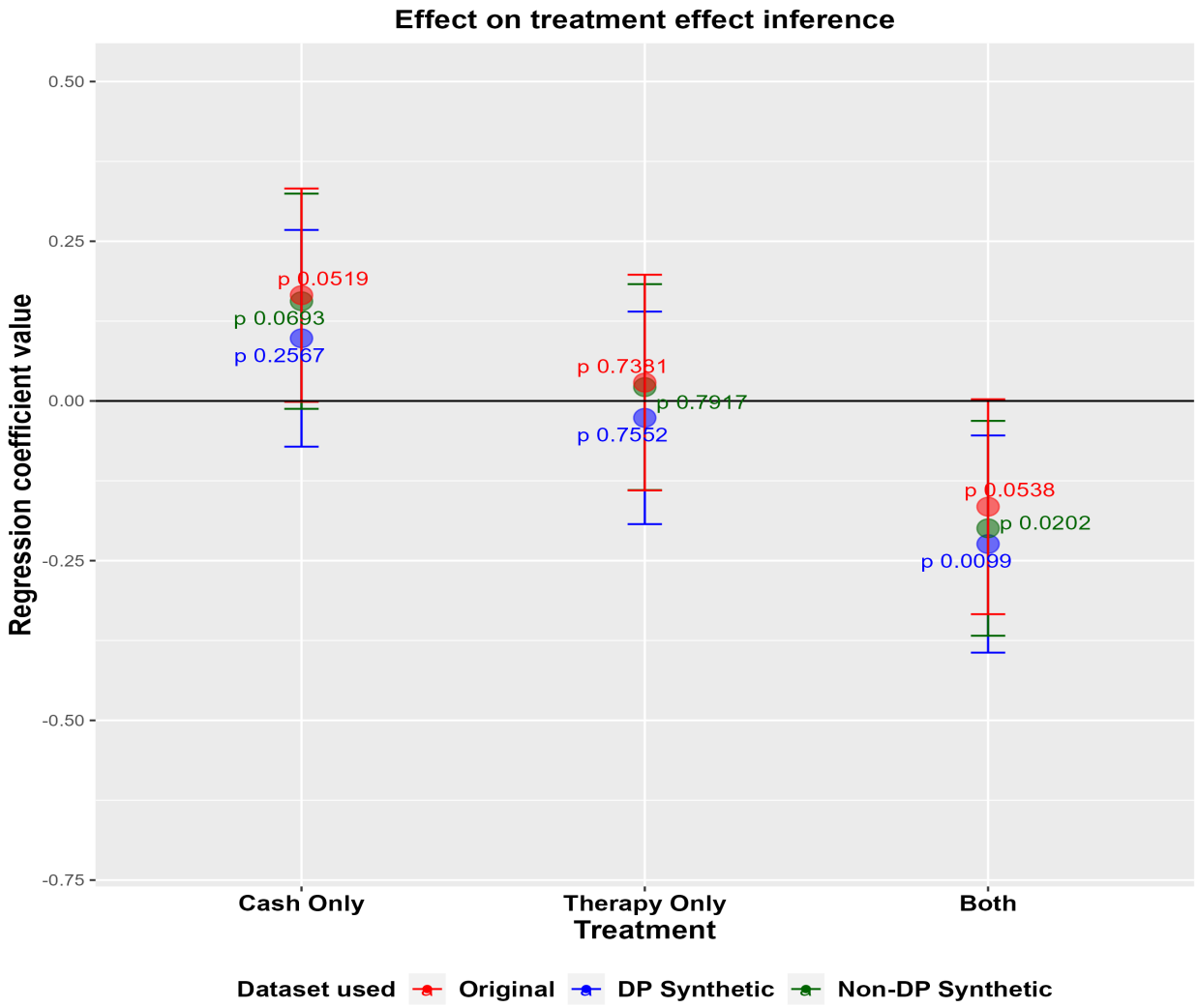

Figure 1: Comparison of inference regarding treatment effect in the Liberia study using the original dataset 1 synthetic dataset generated using privacy budget $\epsilon = 1$ and 1 Non-DP synthetic dataset. Colors red, blue and green correspond to results obtained using original dataset, DP synthetic dataset (with $\epsilon = 1$) and Non-DP synthetic dataset, respectively. The dots correspond to the OLS point estimates of the treatment effects. The intervals correspond to OLS 95% confidence interval for the treatment effects. The p-values for the tests of significance of the individual treatment effects are reported beside the corresponding OLS point estimates.

	Original Dataset	DP Synthetic Dataset	Non-DP Synthetic Dataset
Cash Only	0.10	0.16	0.17
	(0.09)	(0.09)	(0.09)
Therapy Only	−0.03	0.02	0.03
	(0.08)	(0.08)	(0.09)
Both	−0.22*	−0.20*	−0.17
	(0.09)	(0.09)	(0.09)

*$p < 0.05$

Table 12: Comparison of inference regarding treatment effect in the Liberia study using the original dataset, 1 DP synthetic dataset generated using privacy budget $\epsilon = 1$ and 1 Non-DP synthetic dataset. The OLS point estimates of the treatment effects are reported with the corresponding standard errors reported in parentheses under the point estimates. The stars on the OLS estimates indicate whether the p-values for the tests of significance of the treatment effects is less than 0.05 or not.

differentially private data generation procedure, thus empirically proving our expectation that adding differential privacy guarantees is not coming at much extra cost. Further, in Table 17 we compute Metric 5 (MSE) for the sensitive statistic (Variance of Age) based on both DP and non-DP synthetic data generation procedures. Note that, we expect statistics that depend on the covariate data to be potentially distorted due to the privatization procedure implemented and this justifies the large values of MSE for the DP Synthetic generation methods in Table 17.

Variable names	**Metric 1**	**Metric 2**	**Metric 3**	**Metric 4**
(Intercept)	1.00000	0.97000	0.80359	0.02076
Cash Only	1.00000	0.93000	0.79435	0.00816
Therapy Only	1.00000	0.92000	0.77931	0.00781
Both	1.00000	0.94000	0.80048	0.00720
Therapy Block	1.00000	0.98000	0.80903	0.00000
Cash Block	1.00000	0.95000	0.79263	0.00003
Age	1.00000	0.98000	0.73500	0.00001
Barret ASB index	1.00000	0.94000	0.74428	0.00027
Drugs Sell indicator	1.00000	0.95000	0.80788	0.00297
Alcohol self indicator	1.00000	0.96000	0.82439	0.00269
Grass/Opium self indicator	1.00000	0.97000	0.83612	0.00244
Hard Drugs indicator	1.00000	0.98000	0.81889	0.00271
Steal self indicator	1.00000	0.98000	0.82958	0.00254

Table 13: Effect on inference regarding regression coefficients measured using Metrics 1-4 for Liberia study, averaged over 100 independently generated synthetic dataframes (with privacy budget $\epsilon = 0.1$)

Variable names	Metric 1	Metric 2	Metric 3	Metric 4
(Intercept)	1.00000	0.94000	0.80186	0.02202
Cash Only	1.00000	0.93000	0.79505	0.00817
Therapy Only	1.00000	0.92000	0.77917	0.00785
Both	1.00000	0.95000	0.79819	0.00732
Therapy Block	1.00000	0.98000	0.81052	0.00000
Cash Block	1.00000	0.96000	0.79342	0.00003
Age	1.00000	0.95000	0.73386	0.00001
Barret ASB index	1.00000	0.92000	0.74378	0.00028
Drugs Sell indicator	1.00000	0.94000	0.80342	0.00325
Alcohol self indicator	1.00000	0.95000	0.82324	0.00276
Grass/Opium self indicator	1.00000	0.97000	0.83083	0.00257
Hard Drugs indicator	1.00000	0.99000	0.82668	0.00235
Steal self indicator	1.00000	0.96000	0.81604	0.00293

Table 14: Effect on inference regarding regression coefficients measured using Metrics 1-4 for Liberia study, averaged over 100 independently generated synthetic dataframes (with privacy budget $\epsilon = 0.5$)

Variable names	Metric 1	Metric 2	Metric 3	Metric 4
(Intercept)	1.00000	0.96000	0.80359	0.02092
Cash Only	1.00000	0.93000	0.79685	0.00801
Therapy Only	1.00000	0.93000	0.77940	0.00775
Both	1.00000	0.94000	0.79878	0.00722
Therapy Block	1.00000	0.98000	0.80874	0.00000
Cash Block	1.00000	0.96000	0.79245	0.00003
Age	1.00000	0.96000	0.73490	0.00001
Barret ASB index	1.00000	0.94000	0.74657	0.00027
Drugs Sell indicator	1.00000	0.95000	0.80468	0.00309
Alcohol self indicator	1.00000	0.96000	0.82834	0.00261
Grass/Opium self indicator	1.00000	0.98000	0.82588	0.00270
Hard Drugs indicator	1.00000	0.96000	0.81769	0.00264
Steal self indicator	1.00000	0.96000	0.81300	0.00306

Table 15: Effect on inference regarding regression coefficients measured using Metrics 1-4 for Liberia study, averaged over 100 independently generated synthetic dataframes (with privacy budget $\epsilon = 1$)

Variable names	Metric 1	Metric 2	Metric 3	Metric 4
(Intercept)	1.00000	0.95000	0.80074	0.03692
Cash Only	1.00000	0.98000	0.81815	0.00567
Therapy Only	1.00000	0.96000	0.83278	0.00478
Both	1.00000	0.94000	0.80123	0.00676
Therapy Block	1.00000	0.99000	0.82447	0.00000
Cash Block	1.00000	0.96000	0.77586	0.00003
Age	1.00000	0.94000	0.79168	0.00004
Barret ASB index	1.00000	0.93000	0.79746	0.00103
Drugs Sell indicator	1.00000	0.97000	0.80715	0.00618
Alcohol self indicator	1.00000	0.97000	0.79696	0.00447
Grass/Opium self indicator	1.00000	0.92000	0.79709	0.00470
Hard Drugs indicator	1.00000	0.97000	0.78692	0.00639
Steal self indicator	1.00000	0.95000	0.78054	0.00531

Table 16: Effect on inference regarding regression coefficients measured using Metrics 1-4 for Liberia study, averaged over 100 independently generated non-DP synthetic dataframes.

Privacy Budget	Epsilon 0.1	Epsilon 0.5	Epsilon 1	Non-DP Synthesis
MSE of Variance of Age	4481.74	4508.79	4503.16	0.9

Table 17: Effect on value of sensitive statistic (based on covariate data) measured using Metric 5 (MSE) for Liberia study, with varying privacy budget ϵ and non-DP synthesis, averaged over 100 simulations of the sensitive dataframe, using 20 independently generated synthetic dataframes for each sensitive dataframe.

7 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.

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.

Blattman, Christopher, Julian C. Jamison, and Margaret Sheridan. 2017. "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. n.d.. "Replication data for: Reducing Crime and Violence: Experimental Evidence from Cognitive Behavioral Therapy in Liberia." 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).

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. 2006. "Calibrating Noise to Sensitivity in Private Data Analysis." Vol. Vol. 3876, 265–284. https://doi.org/10.1007/11681878_14.

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.

Karr, A. F, C. N Kohnen, A Oganian, J. P Reiter, and A. P Sanil. 2006. "A Framework for Evaluating the Utility of Data Altered to Protect Confidentiality." *The American Statistician*, 60(3): 224–232. https://doi.org/10.1198/000313006X124640.

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.

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.

Roth, Jonathan. 2022. "Pretest with Caution: Event-Study Estimates after Testing for Parallel Trends." *American Economic Review: Insights*, 4(3): 305–22. https://doi.org/10.1257/aeri.20210236.

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.

Wasserman, Larry, and Shuheng Zhou. 2008. "A statistical framework for differential privacy." https://doi.org/10.48550/ARXIV.0811.2501.

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.

8 Acknowledgement and Disclosure of funding

This work is supported by Digital Credit Observatory (CEGA, University of California, Berkeley/Bill and Melinda Gates Foundation (MP)) (PI: Aleksandra Slavković). Soumya Mukherjee and Aratrika Mustafi were supported through NSF award #1702760 (PI: Daniel Kifer). We declare no known conflicts of interest.

Appendix A Results for Simulation Study 1 using beta covariate

Variable names	Metric 1	Metric 2	Metric 3	Metric 4
(Intercept)	1.00000	0.95550	0.78890	0.02966
x1	1.00000	0.94650	0.79535	0.02050
x2	1.00000	0.94900	0.78379	0.07742

Table 18: Effect on inference regarding regression coefficients measured using Metrics 1-4 for Simulation Study 1 with beta covariate, averaged over 100 simulations of the sensitive dataframe, using 20 independently generated synthetic dataframes (with privacy budget $\epsilon = 0.1$) for each sensitive dataframe.

Variable names	Metric 1	Metric 2	Metric 3	Metric 4
(Intercept)	1.00000	0.94450	0.78663	0.02816
x1	1.00000	0.94400	0.79533	0.02049
x2	1.00000	0.94950	0.78748	0.06664

Table 19: Effect on inference regarding regression coefficients measured using Metrics 1-4 for Simulation Study 1 with beta covariate, averaged over 100 simulations of the sensitive dataframe, using 20 independently generated synthetic dataframes (with privacy budget $\epsilon = 0.5$) for each sensitive dataframe.

Variable names	Metric 1	Metric 2	Metric 3	Metric 4
(Intercept)	1.00000	0.95150	0.79378	0.02568
x1	1.00000	0.94900	0.79591	0.02031
x2	1.00000	0.95100	0.79527	0.06172

Table 20: Effect on inference regarding regression coefficients measured using Metrics 1-4 for Simulation Study 1 with beta covariate, averaged over 100 simulations of the sensitive dataframe, using 20 independently generated synthetic dataframes (with privacy budget $\epsilon = 1$) for each sensitive dataframe.

Variable names	Metric 1	Metric 2	Metric 3	Metric 4
(Intercept)	1.00000	0.94700	0.79500	0.02602
x1	1.00000	0.93850	0.79140	0.02180
x2	1.00000	0.94650	0.79965	0.06112

Table 21: Effect on inference regarding regression coefficients measured using Metrics 1-4 for Simulation Study 1 with beta covariate, averaged over 100 simulations of the sensitive dataframe, using 20 independently generated Non-DP synthetic dataframes for each sensitive dataframe.

Privacy Budget	Epsilon 0.1	Epsilon 0.5	Epsilon 1	Non-DP Synthesis
MSE of Variance of x2	0.00069	0.000233	0.000127	0.000059

Table 22: Effect on value of sensitive statistic (based on covariate data) measured using Metric 5 (MSE) for Simulation Study 1 using beta covariate. Results are reported for DP synthesis with varying privacy budget ϵ and non-DP synthesis, each type of synthesis being averaged over 100 simulations of the sensitive dataframe, using 20 independently generated synthetic dataframes for each sensitive dataframe.