IMPROVING PRIVACY AND UTILITY IN AGGREGATE DATA: A HYBRID APPROACH USING POOLDIV

Ivy Min-Zhang
Informatics Laboratory
Nantong University. Nantong, China
minzhangwork1@gmail.com

Kai Chen
Informatics Laboratory
Nantong University. Nantong, China
kai5chen@gmail.com

Wei Percy
Changchun University of Technology
Changchun, China
stariate@changchun.edu

May 5, 2024

ABSTRACT

The increasing need to protect individual privacy in data releases has led to significant advancements in privacy-preserving technologies. Differential Privacy (DP) offers robust privacy guarantees but often at the expense of data utility. On the other hand, data pooling, while improving utility, lacks formal privacy assurances. Our study introduces a novel hybrid method, termed PoolDiv, which combines differential privacy with data pooling to enhance both privacy guarantees and data utility. Through extensive simulations and real data analysis, we assess the performance of synthetic datasets generated via traditional DP methods, data pooling, and our proposed PoolDiv method, demonstrating the advantages of our hybrid approach in maintaining data utility while ensuring privacy.

***K*eywords** differential privacy · pooled regression · estimation and PoolDiv

1 Introduction

Privacy-preserving synthetic data generation and analysis has gained considerable attention in various fields ranging from heath care to social media [1, 2]. The recent surge in the amount of data collected, in health-care for example, provides interesting avenues for exploration and insightful discovery. However, collecting and sharing sensitive, usable, micro-data without disclosing personal information is challenging. In some cases, e.g. through adversarial attacks such as linkage re-identification of individual participants is a possibility.[3]

Differential privacy provides a formal guarantee for confidentially [4, 5]. The mechanism promises that an individual participant's record would remain private even if the adversary has complete knowledge of the rest of the database. This stringent promise has made DP the gold standard for data release [6]. Most agencies provide differentially private synthetic datasets which are often labeled as true representations of the sensitive dataset. [7] Consequently, users must implicitly assume the agency (or curator) has a detailed knowledge of the data characteristics, and that the synthetic data generation model employed by the DP mechanism is well defined. In many cases, users are unable to determine if and how much their analysis results have been impacted by the synthesis process. Inevitably, the accuracy of some analyses deteriorates significantly due to imperfect data generation models. Another formal complain DP synthetic data, is that the mechanism destroys potential insightful data structure such as voids or manifolds.

Another synthetic data generation method that has gained prominence over the years is data aggregation (also referred to as specimen pooling) [6, 8, 9]. Pooling offers an attractive alternative to the differential privacy mechanism as it is much easier to implement in practice. However, it doesn't provide a formal privacy guarantee. The technique randomly combines information from individuals of the same outcome category or exposure of interest, which is then

shared with an analyst. This method suggests preserving privacy by sharing aggregate data instead of individual-level data.[10, 11, 12]

The present study was structured around three primary objectives:

1. **Performance Comparison:** Our first goal was to assess the performance of synthetic data generated from traditional Differential Privacy (DP) mechanisms in comparison to those generated through pooling mechanisms. This evaluation utilized regression modeling to analyze the impact of each method on data utility and accuracy.
2. **Hybrid Mechanism Development:** We proposed a novel hybrid mechanism that combines differential privacy with data pooling, referred to as the pooled-DP mechanism. This initiative was undertaken to enhance privacy protection in pooled data setups without compromising data utility. Subsequently, the performance of this hybrid mechanism was compared with traditional DP and pooling methods to ascertain its effectiveness.
3. **Data Clustering Analysis:** The third objective focused on examining the clustering patterns of synthetic data produced using both the newly developed pooled-DP mechanism and traditional methods. Analyzing these patterns aids in understanding how various synthetic data generation techniques influence the underlying structures and relationships within the data.

These objectives were designed to collectively advance our understanding and refinement of data privacy techniques, ensuring robust privacy protections while maintaining the synthetic data's utility and quality for analytical purposes.

2 Related Work

This section explores the foundational techniques employed in the generation of synthetic data, particularly focusing on ensuring confidentiality and utility. Differential privacy and data pooling methods have been extensively used to address these dual objectives. Here, we delve into differential privacy, its definitions, key mechanisms, and their application in data protection.

2.1 Differential Privacy

Differential Privacy (DP) was introduced as a robust framework for protecting the confidentiality of individual data in datasets used for research and analysis [5]. It is designed to offer strong protection against adversaries who may have access to auxiliary information, thus ensuring that the participation of any individual in a dataset does not significantly influence the outcome of any analysis. For a survey of differential privacy, we refer reader to these articles [13, 14].

Definition 1 (ϵ-Differential Privacy): A mechanism $\mathcal{M}$ is said to satisfy ϵ-differential privacy if for any two adjacent datasets D and D' that differ by a single individual's data, and for all events Z in the output space of $\mathcal{M}$, the probability that $\mathcal{M}$ outputs Z satisfies the following inequality:

$$\log\left(\frac{\Pr[\mathcal{M}(D) = Z]}{\Pr[\mathcal{M}(D') = Z]}\right) \leq \epsilon, \tag{1}$$

where $\epsilon > 0$ is a small constant that determines the level of privacy. The smaller the ϵ, the greater the privacy protection, as the outputs from D and D' are made statistically more indistinguishable.

The definition can be extended to (ϵ, δ)-differential privacy to allow a small probability δ of the mechanism failing to meet the ϵ-differential privacy condition. This relaxation is particularly useful when dealing with complex data or when aiming to improve the utility of the data after applying privacy-preserving techniques.

Within this framework, several mechanisms have been developed to enforce differential privacy:

1. *Randomized Response Mechanism*: This mechanism enhances privacy by introducing randomness in the responses. For example, in a survey, respondents might flip a biased coin in private and only provide their true answer if the coin comes up heads. This simple approach provides a foundational layer of privacy by dissociating the individual's response from their actual data.
2. *Laplace Mechanism*: One of the most common methods for implementing differential privacy involves adding noise generated from a Laplace distribution to the query results. The scale of the Laplace noise is proportional to the sensitivity of the function being computed over the data (denoted as $\mathbb{S}_{\mathcal{M}}$) and inversely proportional to ϵ, ensuring that the added noise maintains the utility of the data while protecting individual privacy.

$$\mathcal{O}(D) = \mathcal{M}(D) + \gamma, \quad \gamma \sim \text{Laplace}\left(0, \frac{\mathbb{S}_{\mathcal{M}}}{\epsilon}\right)$$

3. *Exponential Mechanism*: Used primarily for selecting outputs from a set of possible outcomes. This mechanism assigns probabilities to these outcomes based on a scoring function, which quantifies the utility of each outcome. The selection is skewed towards outcomes with higher utility, adjusted exponentially in accordance with the privacy parameter ϵ.

Definition 2 (Sensitivity): The sensitivity of a query function $\mathcal{M}$, essential for calculating the requisite noise addition in differential privacy, is defined as the maximum change in the output of $\mathcal{M}$ when any single individual's data is changed or removed.

$$\mathbb{S}_{\mathcal{M}} = \max_{(D,D')} \|\mathcal{M}(D) - \mathcal{M}(D')\| \tag{2}$$

where $\|\cdot\|$ denotes the L_1 norm. The effectiveness of these mechanisms is illustrated in Appendix A (Figure 6), showing how different privacy budgets impact the degree of noise and hence the privacy-utility trade-off.

2.1.1 ϵ-Differential Private Synthetic Data

Differential privacy has emerged as a robust framework for data sharing, particularly attractive for its stringent privacy guarantees. This framework has spurred the development of both interactive and non-interactive methods for data release, which adapt various techniques to balance privacy with data utility [15, 16].

We focus on two primary categories of differential privacy mechanisms for generating synthetic data: non-parametric and parametric methods.

Non-parametric Methods: Non-parametric methods do not assume any underlying statistical model for data generation. Instead, they directly utilize the empirical distributions of data attributes to generate synthetic data [5]. One common approach is the histogram perturbation method, where noise is added to the histograms of the data attributes to protect privacy before generating synthetic data from the perturbed histograms.

Algorithm 1: Histogram Perturbation Method

Input : Private database $D = (d_1, d_2, \ldots, d_k)$ of $\mathbb{Z}^+$ and a privacy budget ϵ.

Standardize the entries and add Laplace perturbation:

$$\bar{d}_i = \frac{d_i}{\sum_{i=1}^{k} d_i} + \gamma_i \quad \text{for all } i,$$

where γ_i are iid sampled from Laplace $\left(0, \frac{\mathbb{S}_{\mathcal{M}}}{\epsilon}\right)$ and $\mathbb{S}_{\mathcal{M}} = \max_{(D,D')} \|\mathcal{M}(D) - \mathcal{M}(D')\|_1$.

for $i = 1$ **to** k **do**

$$\bar{d}'_i = \max(\bar{d}_i, 0) \quad \text{and renormalize} \quad \tilde{d}_i = \frac{\bar{d}'_i}{\sum_{j=1}^{k} \bar{d}'_j}$$

end for

return $\tilde{D} = (\tilde{d}_1, \tilde{d}_2, \ldots, \tilde{d}_k)$, *the synthetic data*

Parametric Methods: Parametric methods assume a statistical model for the original data and use parameters estimated from the data to generate synthetic datasets. These methods often use a Bayesian approach, where data is generated from a distribution defined by posterior parameters perturbed according to differential privacy requirements.

Algorithm 2: Multinomial-Dirichlet Synthetic Data Generation

Input : Private dataset $D = (d_1, d_2, \ldots, d_k)$ and the privacy budget ϵ.

Set the prior parameters:

$$\alpha_i = \frac{\tilde{n}}{\exp(\epsilon) - 1} \quad \text{for } i = 1, 2, \ldots, k,$$

Sample the posterior distribution:

$$\tilde{\pi} \sim \text{Dirichlet}(\alpha + D),$$

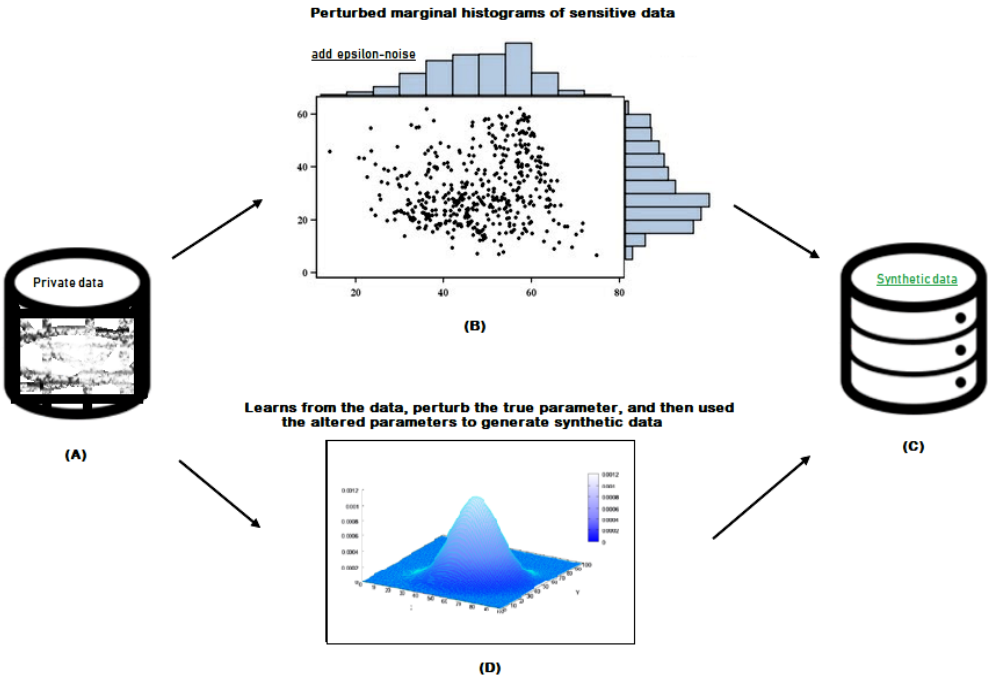
Sample a synthetic dataset:

$$\tilde{D} \sim \text{Multinomial}(\tilde{n}, \tilde{\pi}).$$

return $\tilde{D}$, *differentially private synthetic data of size* $\tilde{n}$

Figure 1 presents a schematic representation of these mechanisms, illustrating the general approach to generating synthetic data under differential privacy. Both methods are vital in scenarios where the original data must remain confidential, yet the utility of the data for analysis cannot be compromised.

Figure 1: Synthetic data generating architecture



These mechanisms facilitate a range of applications from academic research to industry analytics, ensuring that data privacy does not hinder the potential for data-driven insights.

2.1.2 Pooled synthetic data generation

Pooling promises data privacy with minimal loss of data utility. The method, however, doesn't provide a formal privacy guarantee nor does it provide a mechanism for quantifying privacy loss when generating synthetic data with the pooling mechanism. More precisely, pooling doesn't proclaim a guarantee that an individual's data could not be inferred by an adversary. The cost-effectiveness of pooled analysis and the individual data mask it provides has made it appealing among statisticians and epidemiologists who are able to share aggregate data instead of individual-level data. [8, 6, 9]

Various methods have been proposed for pooling based on a relevant outcome of interest or exposure group. We present below a generic algorithm for generating pooled synthetic data.

Algorithm 3: Pooled Synthetic Data Generation

Input : Private dataset $D = (d_1, d_2, \ldots, d_k) \in \mathbb{R}^{n \times k}$ with k attributes and n individual entries.

1 Set the pool size g for the database.

2 Sample the n individuals to create n/g pools.

3 Aggregate individuals in the same pool such that

$$\tilde{d_i} = \sum_{j=1}^{n/g} d_{ji} \quad \text{for all } i$$

return $\tilde{D} = (\tilde{d}_1, \ldots, \tilde{d}_k)$, *the synthetic data* $\in \mathbb{R}^{(n/g) \times k}$ *where* g *is the pool size.*

3 Pooled (ϵ, δ)-Differentially Private Data: PoolDiv

In this section, we introduce PoolDiv, a novel hybrid algorithm that synergistically combines the robust privacy measures of differential privacy from Section 2.1 with the utility-enhancing features of data pooling as described in Section **??**. This approach moderates the stringent privacy parameters typically associated with differential privacy to adopt a more flexible (ϵ, δ)-differential privacy model. This relaxation allows for a practical balance between privacy protection and data utility, paving the way for more effective data analysis in sensitive domains.

The main innovation in PoolDiv lies in its two-stage process where differentially private data are first generated with a relaxed (ϵ, δ) privacy budget, and subsequently, the data are pooled according to predefined group sizes. This method not only enhances privacy but also reduces computational complexity by minimizing the overhead associated with strict privacy controls. The following algorithm describes the detailed steps involved in generating pooled differentially private data using PoolDiv:

Algorithm 4: Generation of Pooled (ϵ, δ)-Differentially Private Data using PoolDiv

Input : Private database $D = (d_1, d_2, \ldots, d_k) \in \mathbb{R}^{n \times k}$ with k attributes and n individual entries, aiming to return a pooled (ϵ, δ)-differentially private dataset $\tilde{D} \in \mathbb{R}^{(n/g) \times k}$ where g is the pool size.

1 Generate differentially private data by applying Algorithm 1 with a relaxed (ϵ, δ) budget.
2 Set the pool size g, and create n/g pools by randomly assigning individuals to each pool.
3 Aggregate the differentially private data within each pool to form $\tilde{D} = (\tilde{d}_1, \ldots, \tilde{d}_k)$.
4 **return** $\tilde{D}$*, the pooled* (ϵ, δ)*-differentially private synthetic dataset*

PoolDiv ensures privacy by making the information about any individual indistinguishable from others within the same pool. This is achieved through the differential privacy guarantees applied prior to pooling, which are quantified by the relaxed (ϵ, δ) parameters. These parameters are chosen based on the desired level of privacy and the specific requirements of the application context. By adjusting the pooling granularity (i.e., the size of g), we can further optimize the balance between data utility and privacy. We hypothesize that PoolDiv, by integrating relaxed differential privacy with pooling, will not only safeguard privacy but also enhance the utility of the synthesized datasets, thereby facilitating more effective and efficient data analysis.

4 Regression on Synthetic Databases

Regression analysis is a powerful statistical tool used to model the relationship between a dependent variable and one or more independent variables. In the context of synthetic databases, this technique helps us understand how well the synthetic data can replicate the relationships present in the original data [17] or if bias within the data might have an impact of statistical modeling [18, 19]. Let us consider a regression scenario where $y = (y_1, \ldots, y_n)^T \in \mathbb{R}^n$ represents the measured responses for n individuals, and $X \in \mathbb{R}^{n \times p}$ is the non-random design matrix of predictors. In this matrix, $x_{i1} = 1$ for each $i = 1, \ldots, n$, to incorporate the intercept term in the regression model.

The model assumes that y is a realization of the linear relationship:

$$Y = X\beta^* + \epsilon,$$

where $\beta^* = (\beta_1^*, \beta_2^*, \ldots, \beta_p^*) \in \mathbb{R}^p$ represents the vector of true regression coefficients, and $\epsilon = (\epsilon_1, \ldots, \epsilon_n)^T$ is the vector of error terms. These error terms are assumed to be independently and identically distributed (iid) with a mean of zero ($\mathrm{E}[\epsilon_i] = 0$) and a constant variance ($\mathrm{Var}[\epsilon_i] = \sigma_*^2$). This assumption of homoscedasticity (constant variance) is crucial for the standard least squares estimation to provide the best linear unbiased estimates (BLUE).

To estimate the regression coefficients from the synthetic data, we employ the least squares criterion, which is formulated as follows:

$$\hat{\beta}^* = \arg\min_{\beta^*} \|Y - X\beta^*\|^2, \tag{3}$$

where $\hat{\beta}^*$ represents the estimated coefficients obtained from minimizing the sum of squared residuals between the observed outcomes and those predicted by the model. This vector of estimates includes $\hat{\beta}_1$, the intercept, and $\hat{\beta}_{-1}$, the coefficients associated with the predictors other than the intercept.

In synthetic data analysis, it is essential to compare the estimated coefficients $\hat{\beta}^*$ derived from the synthetic data to those obtained from the original data. Such comparisons are crucial for validating the quality and utility of the synthetic data, particularly how well it preserves statistical properties like means, variances, and relationships between variables. Since the intercept can vary significantly with scaling transformations of the dataset, our analysis focuses primarily on

$\hat{\beta}_{-1}$, the coefficients of the predictors, which are less sensitive to such transformations. This approach allows for a more stable and meaningful assessment of the synthetic data's fidelity to the original data.

The performance of the regression model on synthetic data can be further evaluated by computing various goodness-of-fit measures such as the R-squared value, Root Mean Square Error (RMSE), and Mean Absolute Error (MAE). These metrics provide insight into how closely the synthetic data approximates the real data's underlying structure and variability, thus indicating the practical utility of the synthetic data generation process in preserving key statistical characteristics.

4.1 Simulation Study

To evaluate the performance of our proposed mechanisms, we conducted a comprehensive simulation study. We generated three covariates (X_1, X_2, X_3) from Normal distributions, each pair having a correlation coefficient of $\rho_{1,2} = \rho_{2,3} = \rho_{1,3} = 0.15$. This setup models realistic scenarios where variables are not entirely independent, which is common in many fields such as economics, social sciences, and biostatistics.

We then constructed the outcome variable y using a linear additive model:

$$y = \beta_1^* X_1 + \beta_2^* X_2 + \beta_3^* X_3,$$

where $\beta^* = \{\beta_1^*, \beta_2^*, \beta_3^*\}$ are the true regression coefficients set to known values for the purpose of the simulation. This model helps to understand how well synthetic data can preserve the relationships inherent in the original data when subjected to privacy-preserving algorithms.

Table 1: Regression coefficients and confidence intervals for private and synthetic data

Predictors	Private data		Randomised		Laplace		Pool 2		Pool 4		Pool 6	
	Est.	CI	Est.	CI	Est.	CI	Est.	CI	Est.	CI	Est.	CI
β_0	0.00	(-0.03, 0.03)	0.00	(-0.09, 0.09)	0.06	(-0.02, 0.13)	-0.00	(-0.06, 0.06)	-0.00	(-0.12, 0.12)	-0.01	(-0.19, 0.18)
X1	0.46	(0.43, 0.49)	0.37	(0.27, 0.46)	0.30	(0.24, 0.36)	0.49	(0.44, 0.53)	0.45	(0.38, 0.51)	0.50	(0.43, 0.57)
X2	-0.13	(-0.16, -0.10)	-0.15	(-0.24, -0.06)	-0.08	(-0.14, -0.02)	-0.12	(-0.16, -0.08)	-0.08	(-0.15, -0.02)	-0.07	(-0.15, 0.00)
X3	-1.68	(-1.71, -1.65)	-1.39	(-1.47, -1.30)	-1.17	(-1.23, -1.11)	-1.67	(-1.71, -1.63)	-1.63	(-1.70, -1.57)	-1.62	(-1.69, -1.54)
Obs	1000		1000		1000		500		250		166	
R^2/R^2adj	0.930/0.930		0.504/0.503		0.590/0.589		0.932/0.932		0.923/0.922		0.928/0.926	

The estimates of β^* using synthetic data generated by our differential privacy mechanisms (histogram perturbation, multinomial-Dirichlet synthesis, and data pooling) are presented in Table 1. We assess how closely these estimates match the true coefficients, which serves as a measure of data utility post-synthesis.

Following this, we used our hybrid algorithm, referred to as PoolDiv (Algorithm 4), to generate (ϵ, δ)-differentially private synthetic data. We experimented with different pool sizes $g = (2, 4, 6, 8)$ to explore how the granularity of pooling affects the accuracy of the parameter estimates.

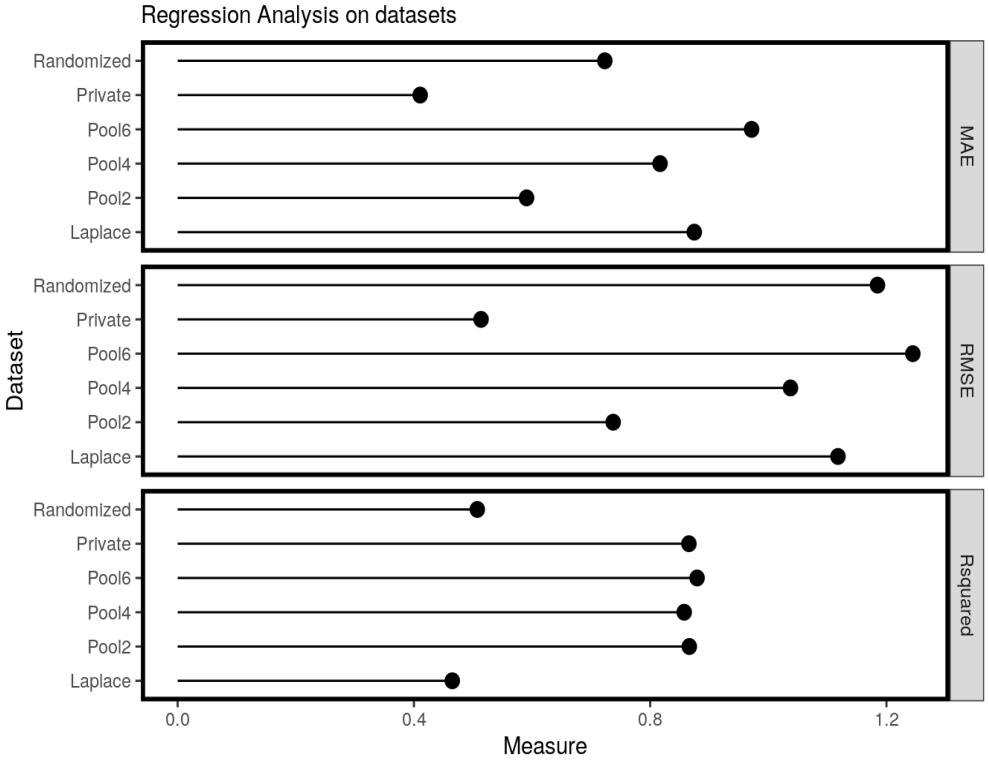
Table 2: Regression Estimates Across Different Pooling Divisions

	Pooled-div 2			Pooled-div 4			Pooled-div 6			Pooled-div 10		
Predictors	Estimates	CI	p	Estimates	CI	p	Estimates	CI	p	Estimates	CI	p
(Intercept)	0.01	(-0.17, 0.18)	0.953	0.01	(-0.34, 0.37)	0.953	0.02	(-0.52, 0.56)	0.943	0.03	(-0.86, 0.91)	0.952
X1	0.34	(0.21, 0.48)	<0.001	0.37	(0.18, 0.55)	<0.001	0.24	(0.02, 0.46)	0.035	0.50	(0.22, 0.78)	0.001
X2	-0.15	(-0.27, -0.03)	0.017	-0.03	(-0.21, 0.16)	0.780	-0.12	(-0.35, 0.12)	0.326	0.16	(-0.15, 0.47)	0.317
X3	-1.46	(-1.58, -1.33)	<0.001	-1.45	(-1.64, -1.27)	<0.001	-1.36	(-1.59, -1.13)	<0.001	-1.24	(-1.53, -0.95)	<0.001
Observations	500			250			166			100		

The effectiveness of the regression models fitted to each dataset was quantitatively evaluated using three metrics: Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and R-squared (R^2). These metrics help in understanding different aspects of model accuracy and fit quality: - **MAE** measures the average magnitude of the errors without considering their direction, serving as a clear indicator of average error magnitude. - **RMSE** provides a measure of error magnitude squared, thus giving higher weight to larger errors. This is particularly useful in emphasizing outliers

or larger deviations from the mean. - $**R^2**$ offers insights into the proportion of variance explained by the model, indicating the strength of the relationship captured by the synthetic data.

Figure 2: Comparison of model performance across different metrics. RMSE is used as the standard criterion for model comparison.



Interpretation: The results from the simulations provide critical insights: - Models using pooled synthetic data consistently yielded closer estimates to the true parameters, suggesting that pooling might help in mitigating the distortion effects introduced by differential privacy noise. For instance, pooling four observations together (g=4) resulted in unbiased parameter estimates, significantly enhancing model accuracy. - On the contrary, increasing the pool size beyond a certain point seemed to deteriorate the quality of estimates, likely due to over-smoothing or loss of critical data variability. - The performance comparison across various synthetic datasets (as shown in Figure 2) highlights that while differentially private datasets generally perform worse than the original data, the introduction of pooling mechanisms tends to improve performance substantially.

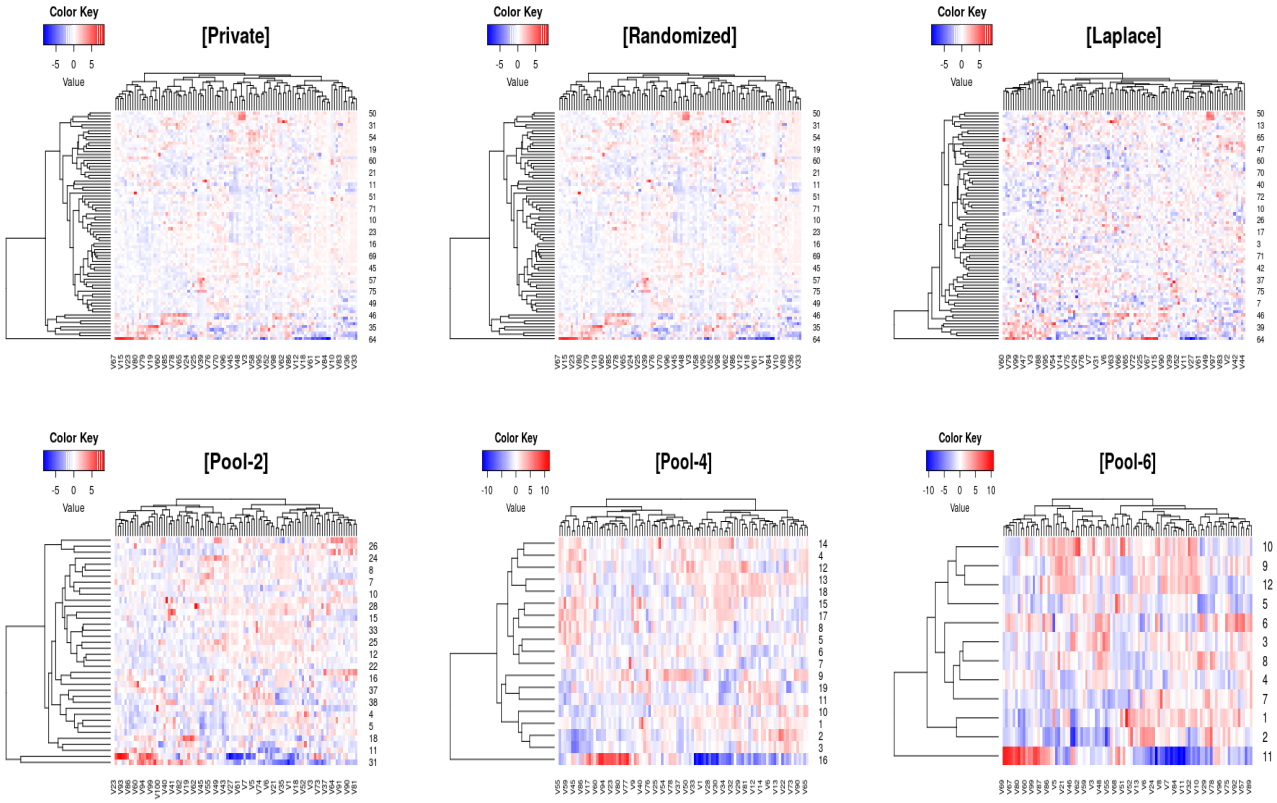
Ultimately, these simulations underscore the importance of choosing appropriate parameters and mechanisms depending on the specific needs of the dataset and the privacy-utility balance required. Our findings suggest that hybrid approaches like PoolDiv can offer a promising compromise, effectively balancing privacy concerns with the need for high-quality synthetic data.

4.2 Case Study

The effectiveness of the algorithms was also tested using a real dataset. We analyzed the lymphoma microRNA data reported by Shipp et al. [20]. This dataset comprises a significant subset of participants diagnosed with non-Hodgkin lymphomas (DLBCL), specifically between 30% and 40% of the study's total population. In total, the study included 58 patients with DLBCL, of whom 32 were successfully cured, while the remaining 26 suffered from fatal or refractory outcomes. The dataset includes measurements of 6,817 gene expression levels. These measurements were used to explore the potential for curative outcomes in patients undergoing a CHOP-based chemotherapy regimen, consisting of cyclophosphamide, adriamycin, vincristine, and prednisone. The analysis aimed to understand how gene expression could influence the effectiveness of this treatment in different patient outcomes.

To assess this high dimension data, we employ a heuristic evaluation of the clustering pattern of the synthetic datasets generated. We present below heatmaps of the synthetic dataset generated. More specifically, we compare the clustering patterns of the private data, randomized outcome data, laplace noise perturbation data, and pooled synthetic datasets.

Figure 3: [Private] clustering of the real dataset studied in [20], [Randomized] Synthetic dataset generated via outcome randomization, [Laplace] via laplace noise perturbation of independently generated histograms, [Pool 2,4,6] Synthetic data generated via pooling samples of sizes 2,4,6 respectively



Interpretation. In the heatmaps presented in Figure 4.2, we see that the true underlying structure is preserved in the pooled dataset and more prominently exhibited as we continue to pool more observations. In the DP mechanism (e.g. Laplace perturbation), we see completely altered noise profiles. This is unsurprising, as DP has been shown to destroy underlying, local data structure.

5 Discussion

Our research has demonstrated that inferential accuracy from traditional differentially private (DP) mechanisms typically falls short compared to pooled analysis. However, our proposed hybrid model, PoolDiv, effectively bridges this gap by combining the robust privacy assurances of DP, specifically under relaxed privacy budgets, with the enhanced utility found in pooled analysis. The performance of the PoolDiv mechanism is on par with traditional DP approaches for low-dimensional data regression and excels in high-dimensional data synthesis.[21, 22, 23]

The PoolDiv mechanism offers significant advantages, particularly in terms of computational efficiency and data structure preservation. Synthetic data generated by PoolDiv tend to be of lower dimensionality, reducing the computational overhead in downstream analyses. For instance, the complexity of estimating regression coefficients in a dataset synthesized by PoolDiv scales as $O\left(\frac{np^2}{g}\right)$, where g represents the pool size, thereby reducing computational cost compared to more complex models. Additionally, PoolDiv excels in maintaining the integrity of the underlying data structure, which is crucial for the validity of subsequent analyses.

Simulation results underscore several key insights: Pooling consistently outperforms traditional DP methods in terms of bias reduction and error rates, as evidenced by lower RMSE scores in pooled synthetic data. Moreover, the hybrid PoolDiv mechanism delivers comparable, if not superior, model fits relative to those achieved using randomized algorithms or Laplace noise perturbation techniques.

For high-dimensional data, standard DP techniques often prove inadequate and present significant challenges. The majority of methods capable of generating usable synthetic datasets, such as those based on neural network models (e.g., see [24, 25]), not only require extensive computational resources but also pose steep learning curves for those not versed in advanced machine learning techniques. In contrast, the PoolDiv approach is not only more straightforward to implement but also effectively preserves essential data characteristics, making it highly beneficial for practical applications.

Despite its strengths, the PoolDiv algorithm is not devoid of limitations. The simplistic nature of the outcome-randomized mechanism it employs, while facilitating plausible deniability, may lead to compounded errors when extensive pooling is applied. This is particularly noticeable when more than two samples are pooled, which can skew the results unfavorably. A potential enhancement could involve integrating Laplace noise perturbation for handling high-dimensional data, although this would need to be carefully balanced to avoid exacerbating the computational complexity.

In summary, while the PoolDiv mechanism represents a significant advancement in synthesizing differential privacy-protected data, continuous improvements and adaptations will be essential to address the evolving challenges in data privacy and synthetic data generation.

References

[1] Pengyue J Lin, Behrokh Samadi, Alan Cipolone, Daniel R Jeske, Sean Cox, Carlos Rendon, Douglas Holt, and Rui Xiao. Development of a synthetic data set generator for building and testing information discovery systems. In *Third International Conference on Information Technology: New Generations (ITNG'06)*, pages 707–712. IEEE, 2006.

[2] H Surendra and HS Mohan. A review of synthetic data generation methods for privacy preserving data publishing. *Int J Sci Technol Res*, 6, 2017.

[3] Charu C Aggarwal and S Yu Philip. *Privacy-preserving data mining: models and algorithms*. Springer Science & Business Media, 2008.

[4] Cynthia Dwork. Differential privacy. *Encyclopedia of Cryptography and Security*, pages 338–340, 2011.

[5] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *Theory of cryptography conference*, pages 265–284. Springer, 2006.

[6] Jeremy A Rassen, Daniel H Solomon, Jeffrey R Curtis, Lisa Herrinton, and Sebastian Schneeweiss. Privacy-maintaining propensity score-based pooling of multiple databases applied to a study of biologics. *Medical care*, 48(6 Suppl):S83, 2010.

[7] Justin Hsu, Marco Gaboardi, Andreas Haeberlen, Sanjeev Khanna, Arjun Narayan, Benjamin C Pierce, and Aaron Roth. Differential privacy: An economic method for choosing epsilon. In *2014 IEEE 27th Computer Security Foundations Symposium*, pages 398–410. IEEE, 2014.

[8] Michael Wolfson, Susan E Wallace, Nicholas Masca, Geoff Rowe, Nuala A Sheehan, Vincent Ferretti, Philippe LaFlamme, Martin D Tobin, John Macleod, Julian Little, et al. Datashield: resolving a conflict in contemporary bioscience—performing a pooled analysis of individual-level data without sharing the data. *International journal of epidemiology*, 39(5):1372–1382, 2010.

[9] P Saha-Chaudhuri and CR Weinberg. Addressing data privacy in matched studies via virtual pooling. *BMC medical research methodology*, 17(1):136, 2017.

[10] Mark A Rothstein, Bartha Maria Knoppers, and Heather L Harrell. Comparative approaches to biobanks and privacy. *The Journal of Law, Medicine & Ethics*, 44(1):161–172, 2016.

[11] Lamin Juwara and Paramita Saha-Chaudhuri. A hybrid covariate microaggregation approach for privacy-preserving logistic regression. *Journal of Survey Statistics and Methodology*, 10(3):568–595, 2022.

[12] Lamin Juwara, Yi Archer Yang, Ana M Velly, and Paramita Saha-Chaudhuri. Privacy-preserving analysis of time-to-event data under nested case-control sampling. *Statistical Methods in Medical Research*, 33(1):96–111, 2024.

[13] Zhanglong Ji, Zachary C Lipton, and Charles Elkan. Differential privacy and machine learning: a survey and review. *arXiv preprint arXiv:1412.7584*, 2014.

[14] Ying Zhao and Jinjun Chen. A survey on differential privacy for unstructured data content. *ACM Computing Surveys (CSUR)*, 54(10s):1–28, 2022.

[15] Frank D McSherry. Privacy integrated queries: an extensible platform for privacy-preserving data analysis. In *Proceedings of the 2009 ACM SIGMOD International Conference on Management of data*, pages 19–30. ACM, 2009.

[16] Arik Friedman and Assaf Schuster. Data mining with differential privacy. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 493–502. ACM, 2010.

[17] TA Reddy and DE Claridge. Using synthetic data to evaluate multiple regression and principal component analyses for statistical modeling of daily building energy consumption. *Energy and buildings*, 21(1):35–44, 1994.

[18] Sahra Ghalebikesabi, Harry Wilde, Jack Jewson, Arnaud Doucet, Sebastian Vollmer, and Chris Holmes. Mitigating statistical bias within differentially private synthetic data. In *Uncertainty in Artificial Intelligence*, pages 696–705. PMLR, 2022.

[19] Lamin Juwara, Alaa El-Hussuna, and Khaled El Emam. An evaluation of synthetic data augmentation for mitigating covariate bias in health data. *Patterns*, 2024.

[20] STEWART SHIPP and SEMIR ZEKI. The functional organization of area v2, ii: the impact of stripes on visual topography. *Visual neuroscience*, 19(2):211–231, 2002.

[21] Eugene Bagdasaryan, Omid Poursaeed, and Vitaly Shmatikov. Differential privacy has disparate impact on model accuracy. *Advances in neural information processing systems*, 32, 2019.

[22] Dana Pessach and Erez Shmueli. A review on fairness in machine learning. *ACM Computing Surveys (CSUR)*, 55(3):1–44, 2022.

[23] Viraj Kulkarni, Milind Kulkarni, and Aniruddha Pant. Survey of personalization techniques for federated learning. In *2020 fourth world conference on smart trends in systems, security and sustainability (WorldS4)*, pages 794–797. IEEE, 2020.

[24] Qian Wang, Yan Zhang, Xiao Lu, Zhibo Wang, Zhan Qin, and Kui Ren. Real-time and spatio-temporal crowd-sourced social network data publishing with differential privacy. *IEEE Transactions on Dependable and Secure Computing*, 15(4):591–606, 2016.

[25] Nazmiye Ceren Abay, Yan Zhou, Murat Kantarcioglu, Bhavani Thuraisingham, and Latanya Sweeney. Privacy preserving synthetic data release using deep learning. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 510–526. Springer, 2018.

6 Appendix A

Figure 4: Differential privacy generating mechanisms for different privacy budgets. [L] Randomized response mechanism, [M] Laplace noise mechanism, [R] Exponential noise.

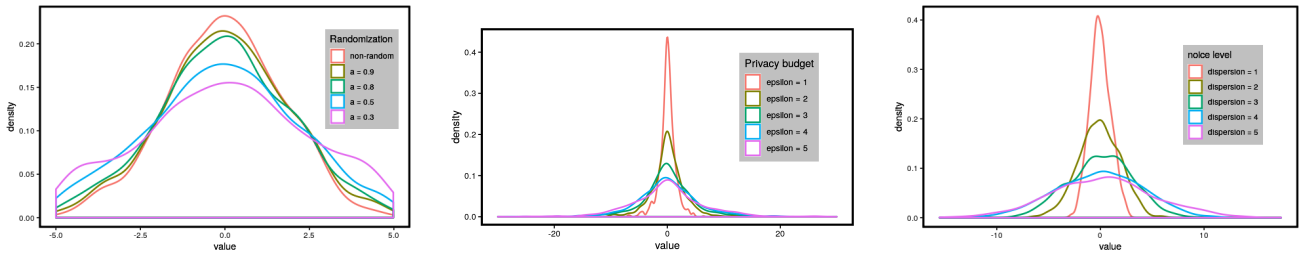
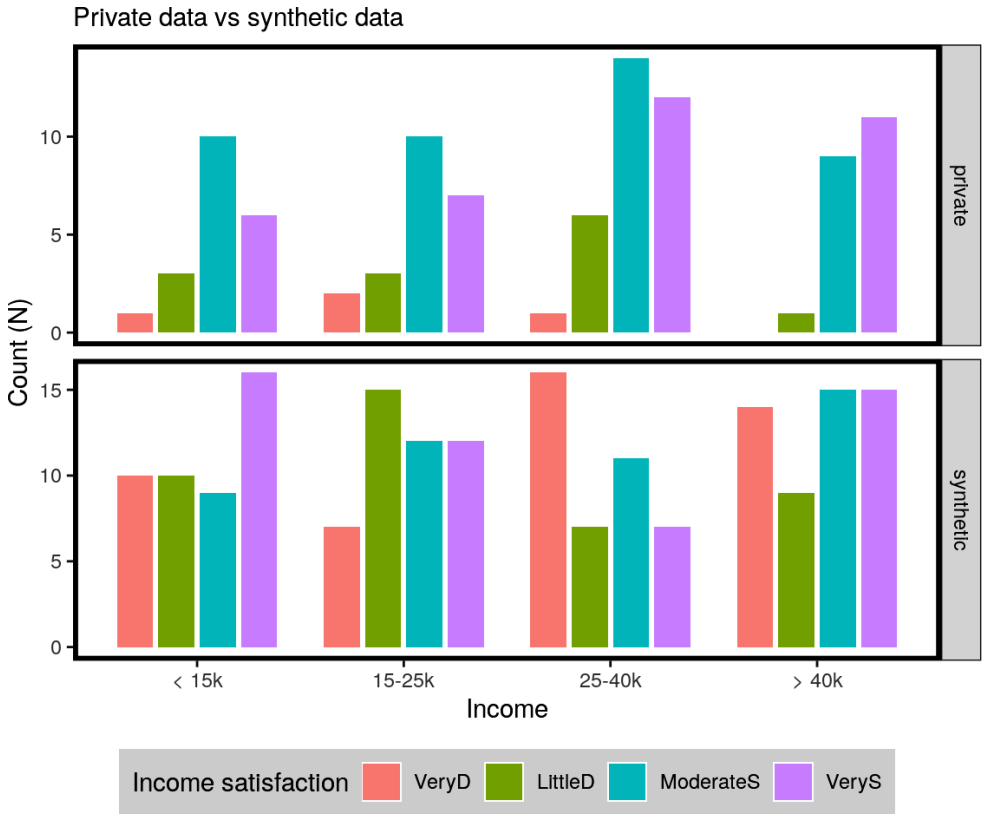



Figure 5: Toy example of private and synthetic data generated using the histogram perturbation algorithm