Performance of Healthcare Analysis Under LDP

Andres Hernandez-Matamoros[1][0000−0002−4896−2909] and Hiroaki Kikuchi[2][0000−0002−0903−8430]

[1] Organization for the Strategic Coordination of Research and Intellectual Property, Meiji University, Tokyo 164-8525, Japan
[2] School of Interdisciplinary Mathematical Sciences, Meiji University, Tokyo 164-8525
https://www.kikn.fms.meiji.ac.jp/
{matamoros,kikn}@meiji.ac.jp

Abstract. This study investigates the efficacy of healthcare data analysis within the stringent framework of local differential privacy (LDP). By ensuring the privacy of individual data points while allowing for comprehensive analysis, LDP presents unique challenges and opportunities. This research compares various methodologies to determine their performance in the context of healthcare data. Our analysis highlights the suitability of different approaches, particularly emphasizing the balance between privacy preservation and data utility. The findings reveal that method Castell which is based on probability matrixes exhibit superior performance. This study thus contributes to the field of privacy-preserving healthcare analytics, offering valuable insights into the application of LDP in maintaining data integrity and utility.

Keywords: Local Differential Privacy · privacy budget · healthcare data.

1 Introduction

The exponential growth of healthcare data, driven by the widespread adoption wearable devices, presents unprecedented opportunities for revolutionizing medicine. Insights derived from these abundant resources have the potential to enhance treatment strategies, improve disease diagnosis, and optimize healthcare resource allocation. However, the transformative potential of this data is contingent upon effectively balancing the need to unlock its value while safeguarding patient privacy. Traditional anonymization techniques have proven insufficient, prompting the adoption of differential privacy (DP) and its extension to local differential privacy (LDP). LDP [7] enhances privacy by anonymizing data at the point of collection, ensuring that even if the data is compromised, individual identities remain protected.

In healthcare data analysis under LDP, estimating joint probability distributions (JPD) is crucial for capturing correlations and dependencies between various health variables. Techniques such as Bloom filters (BF)[4] and randomized response (RR)[5] are commonly used but can face challenges related to memory consumption and dimensionality. To address these issues, Kikuchi et al.

proposed use Castell algorithm [6] which applies probability matrixes to estimate JPD more efficiently. This method improve accuracy compared to traditional BF and RR techniques. This paper evaluates these approaches using open healthcare datasets, demonstrating that Castell approach provides superior performance.

Table 1. Difference between Lopub, Locop, Br, and Castell.

Approach	Lopub [1]	Locop [3]	Br [2]	Castell [6]
Perturbation	Bloom filter			Randomize Response
	Randomize Response			
Estimation	Lasso with Expectation Maximization algorithm	Lasso with Gaussian Copula	Bayesian Ridge Regression	Probability matrixes

The following sections of the paper are organized as follows: Section 2 briefly describes the LDP problem, as well as the methods used for anonymizing patient data and estimating JPD. These methods are employed by various approaches, including Lopub [1], Locop [3], Br [2], and Castell [6]. Following this, section 3 presents the experiments conducted, detailing the datasets used and the metrics employed to evaluate the performance of the approaches. Finally, the paper concludes with the Conclusion section.

2 Preliminaries

In LDP approaches, patients who wish to share their information with a central server must encode and perturb their data before sending it to the server. By doing this, patients preserve their anonymity. We generalize the LDP problem, considering N patients with C attributes as $u_j^i; (i = 1, \ldots, N; j = 1, \ldots, C)$, where the superindex i means the number of user and the subindex j means the number of attribute for the patient n. The domain of each attribute is denoted $\Omega_j = \{\omega_j^1, \ldots, \omega_j^{|\Omega_{j,k}|}\}$. The cardinality $|\Omega_j|$ means the number of elements in the attribute j.

2.1 Local Differential Privacy

LDP [7] offers a strong privacy guarantee that allows patients to trust themselves, rather than relying on a centralized authority.

Definition 1 (Local Differential Privacy). *An algorithm S satisfies ϵ-LPD if, for any two records u and w and any output $\tilde{u}$ within the range of outputs of the algorithm $S(\tilde{u})$, the equation (1) holds:*

$$Pr[S(u) \in \tilde{u}] \leq e^{\epsilon} Pr[S(w) \in \tilde{u}]. \quad (1)$$

2.2 Approaches

These subsections outline the approaches to anonymize patient's data and estimate the JPD of Lopub[1], Locop[3], Br[2], and Castell[6]. For more comprehesive details, please consult the original papers. The goal is to share patient data while preserving their privacy, enabling researchers to use the data to gain demographic insights from the dataset.

2.2.1 Anonymization Algorithms This subsection briefly describes the anonymization algorithms used by the studied approaches. Two main methods are discussed: the Bloom filter and randomized response (BF-RR) approach used by Lopub, Locop, and Br, and the direct randomized response (RR) on raw data used by Castell.

2.2.1.1 ***BF-RR*** The BF-RR algorithm, proposed by Ren et al.[1] and utilized by Lopub, Locop, and Br, involves encoding and perturbing patient data. Each patient record is encoded using Bloom filters (BF)[4] with multiple hash functions, producing bit strings that represent the attributes. This encoded data is then subjected to randomized response (RR)[5] , where each bit of the bit string is randomly flipped with a certain probability to ensure privacy. This process generates a randomized Bloom filter that is transmitted to the server, preserving user privacy through local randomization techniques.

Encoding information of users The user input is represented by the BF $\mathcal{H}_j$ for the j^{th} attribute; the patient uses h hash functions $\mathcal{H}_{j,1}, \ldots, \mathcal{H}_{j,h}$ to map the data into a bit string, in our experiments we set $h = 4$. One variable is used for encoding the information of the users: the maximum number of bits m_j, as $m_j = \frac{ln(1/p)}{(ln2)^2}|\Omega_j|$. Where p is false positive probability, in our experiments we set $p = 0.022$. After encoding step is applied, the user input is represented as $\boldsymbol{y} = (y[1], y[2], y[3], ..., y[m_j])$

Perturbing the data RR allows interviewer to give their answers while keeping confidentiality. Randomly the question is to be answered truthfully or not, unknown to the interviewer. RR response is applied after encoding step, each bit b, $b \in \{1, m_j\}$ of $\boldsymbol{y}$ is randomly flipped. The output $\hat{y}[b]$ is defined as follows:

$$\hat{y}[b] = \begin{cases} y[b] \textit{ with probability of } 1-f, \\ 1 \textit{ with probability of } f/2, \\ 0 \textit{ with probability of } f/2, \end{cases} \tag{2}$$

User privacy is preserved through the assurance of confidentiality provided by individualized local randomization methods, which users autonomously apply to their data entries. The local perturbation of d attributes can achieve ϵ-local differential privacy (ϵ-LDP), with h being the number of hash functions in the Bloom filter and f the flip bit probability, as given by $\epsilon = 2dh \ln \frac{2-f}{f}$.

2.2.1.2 ***RR*** Let u_j^i represent the true value of the j^{th} attribute for the i^{th} patient. Each patient flips a biased coin with probability $p = \frac{e^\epsilon}{e^\epsilon + |\Omega_j| - 1}$, where ϵ is the privacy budget. If the coin shows heads, the patient reports their true value u_j^i. If the coin shows tails, the patient selects a random value $\hat{u}_j^i$ from the set of possible responses Ω_j. Finally, patients report this data to the server.

2.2.2 Estimating Joint Probability Distributions (JPD) This subsection details the algorithms used to estimate JPD. Four methods are highlighted: regression analysis (Lopub, Br), Lasso regression with Gaussian copula (Locop), and Castell.

2.2.2.1 ***Regression Algorithms*** The central server receives the noised information of the patients. Next, it counts the frequency of the perturbed value $\hat{y}[b]$. Then, the original count $y[b]$ is estimated as $y[b] = (\hat{y}[b] - fN/2)/(1-f)$. After the original count is computed, the candidate bit matrix M is $M = [\mathcal{H}_1(\Omega_1) \times \mathcal{H}_2(\Omega_2) \times \cdots \times \mathcal{H}_k(\Omega_k)]$, where $\mathcal{H}_j(\Omega_j)$ is a matrix for $j = 1, \ldots k$, k is the number of attributes for which the JPD is to be estimated. Finally, the coefficients β of the regression algorithm (Lasso[1] for Lopub and Bayesian ridge regression [2] for Br) $y = M\beta$, is used to estimate the JPD as $JPD = \frac{\beta}{\text{sum}(\beta)}$.

2.2.2.2 ***Lasso Regression with Gaussian Copula*** The central server receives the perturbed patient information. Initially, it estimates the one-dimensional and two-dimensional distributions using the methodology outlined by Ren [1]. Subsequently, it calculates the Pearson correlation coefficient matrix for the k-dimensional attributes, which is utilized in modeling the multivariate Gaussian Copula. For any pair of attributes (U_w, U_v) $(w, v = 1, ..., d)$, the Pearson correlation ρ_{U_w, U_v} is computed, forming the correlation matrix R. Finally, the multivariate Gaussian Copula is formulated based on the Gaussian Joint Distribution $\Phi(0, R)$.

2.2.2.3 ***Castell*** After the patients apply RR to their data and transmit it to the central server, the central server has the dataset $\hat{U}$ and proceeds with JPD estimation. It derives empirical values, κ_j, for each attribute j from the randomized patient data. Next, for each j^{th} attribute in $\hat{U}$ a probability matrix Λ_j is defined as:

$$\Lambda_{j_{k,l}} = \begin{cases} \frac{1-p}{|\Omega_j|-1} & \text{if } k \neq l, \\ p & \text{if } k = l, \end{cases} \tag{3}$$

where $p = \frac{e^\epsilon}{e^\epsilon + |\Omega_j| - 1}$, Λ_j is a square matrix of size $|\Omega_j| \times |\Omega_j|$.

The empirical values κ_j, along with the probability matrix λ_j calculated using equation (3), constitute the foundational steps of the algorithm. With iterative attribute processing marked by i, the algorithm computes transformation parameters γ, utilizing attribute-specific probability matriXes λ_i and empirical values κ_i. Through this systematic approach, a modified matrix Γ is constructed,

employing the inverse of the probability matrix Λ_i^{-1} and γ, thus updating the JPD via the product with Γ. The Castell[6] algorithm is outlined in 1.

Algorithm 1 Castell algorithm

attributes $(i = 1, 2, ..., d)$
$JPD = 1$
for *each* i **do**
$\gamma = \Lambda_i \cdot \kappa_i$
$\Gamma = \Lambda_i^{-1} \cdot \gamma$
$JPD' \leftarrow \Gamma \times JPD$
end for

Table 2. Datasets characteristics.

Dataset	# Patients (N)	# Attributes (C)
Skin Cancer [9]	10,015	5
Nursery [8]	12,960	9
Diabetes [10]	70,692	18

3 Experiments

3.1 Datasets

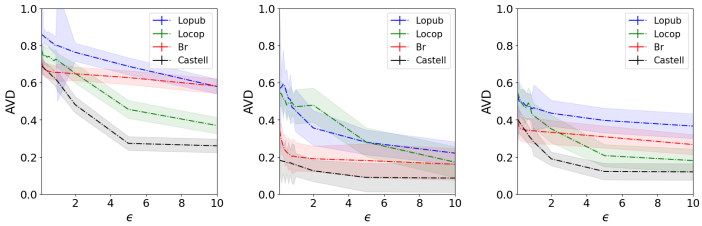
The paper evaluates these approaches using open healthcare datasets, comparing their performance in terms of accuracy. Table 2 presents the statistics of the size of the datasets, including characteristics such as attributes and patients. Through a discretization process, the datasets have been transformed, converting their continuous attributes into five distinct categories. The datasets exhibit significant variations in patient population and attributes. The Diabetes dataset boasts the highest number of patients and the highest number of attributes.

3.2 JPD estimation

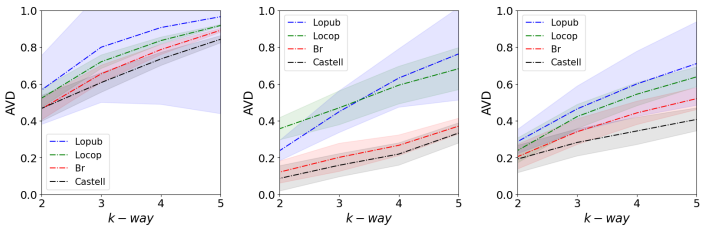
We randomly select a subset of k attributes from each dataset and compute their JPD in a k-way. This process is repeated one hundred times.

To analyze performance when estimating JPD, we use the average variant distance (AVD) metric to quantify the disparity between real and estimated data. AVD, as employed by [1, 3, 2, 6]. It is defined as $AVD = \frac{1}{2}\sum_{\omega \in \Omega} |P(\omega) - Q(\omega)|$, where a value close to zero indicates more accurate JPD. The results are presented in the Figure 1.

Figure 1 presents a comparative analysis of the performance of four LDP approaches across different datasets. The upper part of the figure illustrates the



AVD vs Privacy Budget (ϵ) per attribute with 3-way.



AVD vs k-way, with $\epsilon = 1$.

Skin cancer **Nursery** **Diabetes**

Fig. 1. AVD vs Privacy Budget (ϵ) per attribute with 3-way.

AVD and privacy budget ϵ, while the lower part shows the AVD and k-way comparison. These approaches include the simple Lopub, which relies solely on LASSO regression, Locop, Br, and Castell.

On the Y-axis, AVD indicates the level of distortion caused by anonymization after attempting to recover the Joint Probability Distribution (JPD). Lower AVD values imply less error, which is desirable. In the upper figure, the X-axis represents ϵ, the privacy budget determining the degree of privacy protection offered. Lower ϵ values signify stronger privacy guarantees but may result in higher AVD. On the bottom figures, the X-axis represents the k-way evaluation.

Each line in the graph represents the performance of an LDP approach on a specific dataset, with points depicting the trade-off between AVD and the privacy budget (ϵ). Notably, the Castell approach consistently achieves the lowest AVD across all datasets for a given ϵ value, indicating its effectiveness in minimizing data distortion within a specific privacy budget.

In contrast, the Lopub, Br, and Locop approaches tend to have higher AVD than Castell for the same ϵ value, suggesting greater data distortion to achieve similar privacy levels. The Br approach shows close but not similar performance to Castell for some datasets.

However, it is important to note that algorithm performance can vary depending on the dataset, as shown in the bottom of Figure 1. For example, Castell significantly outperforms Locop, Lopub, and Br (which presents close but not similar values) on the Diabetes dataset, while the difference is less pronounced

on the Skin cancer dataset due to differences in the average absolute Pearson correlation coefficient (AAR) between datasets, as examined by [2]. Overall, Figure 1 shows that Castell is the preferred approach for minimizing AVD while maintaining a low ϵ.

4 Conclusion

In summary, our analysis of four LDP approaches (Lopub, Locop, Br, and Castell) has provided valuable insights into their performance on healthcare datasets. We discovered that the Castell approach consistently achieved the lowest Average Value Distortion (AVD) across various datasets for a given privacy budget, indicating its effectiveness in minimizing data distortion while preserving privacy. Furthermore, we observed that dataset characteristics influenced the performance of LDP approaches. Our findings suggest that the Castell approach is preferable for minimizing AVD while maintaining a low epsilon value. Future research should prioritize comparing memory consumption between approaches and generating synthetic data using joint probability distributions as estimated by Castell. Additionally, evaluating their performance on large medical datasets is essential.

Acknowledgment

This work was supported by JST, CREST Grant Number JPMJCR21M1, Japan.

References

1. Ren, X.; Yu, C.M., Yu, W., Yang, S., Yang, X., McCann, J.A. and Philip, S.Y. LoPub: High-Dimensional Crowdsourced Data Publication with Local Differential Privacy. *IEEE Trans. Inf. Forensics Secur.* 2018, *13*, 2151–2166. https://doi.org/10.1109/TIFS.2018.2812146.
2. Hernandez-Matamoros, Andres, and Hiroaki Kikuchi. 2024. "Comparative Analysis of Local Differential Privacy Schemes in Healthcare Datasets" Applied Sciences 14, no. 7: 2864. https://doi.org/10.3390/app14072864
3. Wang, T.; Yang, X.; Ren, X.; Yu, W.; Yang, S. Locally Private High-Dimensional Crowdsourced Data Release Based on Copula Functions. *IEEE Trans. Serv. Comput.* 2022, *15*, 778–792. https://doi.org/10.1109/TSC.2019.2961092.
4. Bloom, B.H. Space/Time Trade-Offs in Hash Coding with Allowable Errors. *Assoc. Comput. Mach.* 1970, *13*, 422–426.
5. Warner, S.L. Randomized Response: A Survey Technique for Eliminating Evasive Answer Bias. *J. Am. Stat. Assoc.* 1965, *60*, 63–69.
6. Hiroaki Kikuchi, Castell: Scalable Joint Probability Estimation of Multi-dimensional Data Randomized with Local Differential Privacy. 2022, arXiv preprint https://arxiv.org/abs/2212.01627.
7. Kasiviswanathan, S.P.; Lee, H.K.; Nissim, K.; Raskhodnikova, S.; Smith, A. What Can We Learn Privately? In Proceedings of the 49th Annual IEEE Symposium on Foundations of Computer Science, Philadelphia, PA, USA, 25–28 October 2008.

8. Rajkovic, V. Nursery. UCI Machine Learning Repository. https://archive.ics.uci.edu/dataset/76/nursery (accessed on 15 November 2023). 1997.
9. Tschandl, P., Rosendahl, C. and Kittler, H. The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. Sci Data 5, 180161 (2018). https://doi.org/10.1038/sdata.2018.161, https://doi.org/10.7910/DVN/DBW86T.
10. CDC. CDC—2015 BRFSS Survey Data and Documentation. 2015. Available online: https://www.cdc.gov/brfss/annual_data/annual_2015.html.(accessed on 15 November 2023).