Differential Privacy Techniques for Healthcare Data

Rishabh Subramanian
Chirec International School
Hyderabad, India
Email: rishabh.subramanian@gmail.com

2022 International Conference on Intelligent Data Science Technologies and Applications (IDSTA) | 978-1-6654-9960-6/22/$31.00 ©2022 IEEE | DOI: 10.1109/IDSTA55301.2022.9923037

***Abstract*—This paper analyzes techniques to enable differential privacy by adding Laplace noise to healthcare data. First, as healthcare data contain natural constraints for data to take only integral values, we show that drawing only integral values does not provide differential privacy. In contrast, rounding randomly drawn values to the nearest integer provides differential privacy. Second, when a variable is constructed using two other variables, noise must be added to only one of them. Third, if the constructed variable is a fraction, then noise must be added to its constituent private variables, and not to the fraction directly. Fourth, the accuracy of analytics following noise addition increases with the privacy budget, ϵ, and the variance of the independent variable. Finally, the accuracy of analytics following noise addition increases disproportionately with an increase in the privacy budget when the variance of the independent variable is greater. Using actual healthcare data, we provide evidence supporting the two predictions on the accuracy of data analytics. Crucially, to enable accuracy of data analytics with differential privacy, we derive a relationship to extract the slope parameter in the original dataset using the slope parameter in the noisy dataset.**

***Index Terms*—ata privacy, Differential privacy, Healthcareata privacy, Differential privacy, HealthcareD**

I. Introduction

In the wake of the COVID-19 crisis, the healthcare sector has become critical to society. The pandemic has also brought an increased focus on research into subjects that require data from hospitals and other medical agencies. However, publicly sharing this data, even when done with the patient's approval and stripped of personal identifiers, can compromise individuals' privacy. [1] show how researchers could re-identify 95 percent of individual adults from the National Health and Nutrition Examination Survey using machine learning techniques; [2] similarly show a high re-identification rate of data. A prominent example is Governor William Weld, the former Governor of Massachusetts, who was subjected to such re-identification using a linkage attack ([3]). While such re-identification attacks can be countered using methods such as coarsening the representation of key variables ([4]), such techniques may affect the efficacy of healthcare practices and the quality of healthcare research. Therefore, the usage of viable privacy protection methods is cardinal to our increasingly healthcare-dependent society. This paper studies techniques to enable differential privacy to healthcare data.

Differential privacy enables the provision of privacy to individuals in a dataset when such data is shared publicly ([5]). As differential privacy changes the *process* for accessing data, rather than the database itself, it can enable patients' privacy without compromising on the usability of the data to provide accurate diagnoses, prescriptions, etc. Differential privacy can, therefore, provide sufficient privacy to patients.

We examine techniques to enable differential privacy by adding Laplace noise to healthcare data. Our key results are as follows. First, healthcare data contain natural constraints for data to take only integral values. For example, the number of deaths or cases due any given disease must be integral. Thus, we show that drawing only integral values does not provide differential privacy. In contrast, rounding randomly drawn values to the nearest integer does provide differential privacy. Second, when a variable is constructed using two other variables, noise must be added to only one of them. Third, if the constructed variable is a fraction, then noise must be added to its constituent private variables, and not to the fraction directly. Fourth, the accuracy of analytics following noise addition increases with the privacy budget, ϵ, and the variance of the independent variable. Finally, the accuracy of analytics following noise addition increases disproportionately with an increase in the privacy budget when the variance of the independent variable is greater. Using actual healthcare data, we provide evidence supporting the two predictions on the accuracy of data analytics. Most importantly, to enable accuracy of data analytics with differential privacy, we derive a relationship to extract the slope parameter in the original dataset using the slope parameter in the noisy dataset. We then examine the quality of analytics following addition of noise and confirm the predictions using an actual healthcare dataset.

Our study contributes to the literature studying differential privacy in the context of heathcare data ([6], [7], [8], [9], [10], [11], [12], [13], [14], [15], [16], [17], [18], [19], [20], [21], [22], [23], [24], [25], [26], [27]). The key contribution of our study is to analyse techniques for adding Laplace noise to healthcare data so that differential privacy is preserved, and analyse theoretically and empirically the effects of applying these techniques on healthcare data analytics.

The paper is structured as follows. Section II focuses on the applications of differential privacy to healthcare by analyzing the characteristics of healthcare data and devising specific approaches to satisfy differential privacy for each data type and Section III analyzes the effects of noise addition on the accuracy of healthcare analytics, particularly ordinary least squares and difference-in-difference. Section IV concludes the study.

II. Methods for Differential Privacy in Healthcare data

A. Characteristics of Healthcare Data

Healthcare data contains both categorical data and numeric data, either for each individual or aggregated for some community/collection. When aggregated, both categorical and numeric data contain numbers as they are sums or counts of values at the individual level. Therefore, there are infinite values for both aggregate categorical and numeric data. Thus, we can add distortions, such as noise (addressed in section II-B), to aggregate values as a distribution has infinite values that it may provide. At the aggregate level, all the values must also be integral as they are all sums or counts.

Many fields in healthcare datasets are correlated or contain natural constraints. These correlated fields in data may contain values such as per capita, per case measures, etc., that depend on other fields. Similarly, natural constraints may be that the

values of a variable must be integral. Adding distortions to these data can impact the correlations. Researchers may use these correlations as sanity checks to examine the quality of the data. If the independent distortions produce results that do not maintain these correlations, they would erode the trust in the data and, thereby, enact barriers to the acceptability of the data privacy techniques.

B. Noise Addition to Healthcare Data

Adding noise to data is a widely used method to satisfy differential privacy. Random values drawn from a Laplace probability distribution are added them to each value in a dataset to protect an individual's privacy. Noise addition is used mainly for aggregate data, sums or counts of individual level data because adding noise to individual data may change material information about the individual.

1) Adding Noise to Aggregate Categorical Data: **The methods for noise addition differ for categorical data and numeric data. This section discusses the method used for noise addition to categorical data.**

If we take that the databases D and D' have values k and k-1 for some field, taking noise from a Laplace distribution with $\mu = k$ and $b = \frac{1}{\epsilon}$ achieves ϵ-differential privacy. The derivation for this is as follows:

$$Laplace(x|k, \frac{1}{\epsilon}) = \frac{1}{2 \cdot \frac{1}{\epsilon}} \cdot e^{-\frac{|x-k|}{\frac{1}{\epsilon}}}$$

Therefore, the ratio between probability of an output from a dataset with the target and one without the target—having values are k and (k-1) and representing whether or not the target is part of the group—is:

$$\frac{Laplace(x|k, \frac{1}{\epsilon})}{Laplace(x|k-1, \frac{1}{\epsilon})} = e^{-\epsilon \cdot |x-k| - (-\epsilon \cdot |x-(k-1)|)}$$

As the variable is a whole number, $k-1 \leq x < k$ is not possible. So, there are two possible cases, $x < k-1$ and $x \geq k$.

In case 1, where $x < k-1$, we get:

$$\therefore \frac{Laplace(x|k, \frac{1}{\epsilon})}{Laplace(x|k-1, \frac{1}{\epsilon})} = e^{-\epsilon}$$

In case 3, where $x \geq k$, we get:

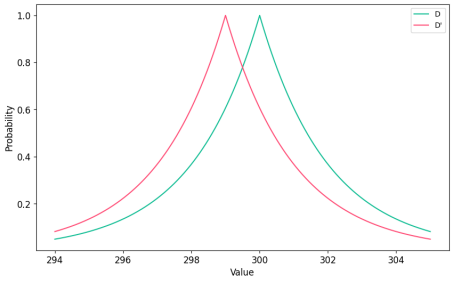
$$\therefore \frac{Laplace(x|k, \frac{1}{\epsilon})}{Laplace(x|k-1, \frac{1}{\epsilon})} = e^{\epsilon}$$

Therefore, we get the following result:

Result 1: Differential privacy is satisfied in all cases when noise is added from a Laplace distribution with $b = \frac{1}{\epsilon}$ to aggregate categorical data.

Figure 1 provides the intuition for the result. If we were to take a point such as 302, the attacker cannot infer what the initial value was as there is a possibility that the initial value was either 300 or 299. Thus, the attacker cannot find any additional information about the target.

Fig. 1: **Noise Addition to Aggregate Categorical Data**



2) Adding Noise to Aggregate Numeric Data: **The method for noise addition discussed in section II-B1 works efficiently for aggregate categorical data, but not necessarily for aggregate numeric data. Consider the number of patients a hospital received on a particular day. Say the maximum number of times a patient visited the hospital on that day was five. If there are multiple entries for one patient, the effect of removing one patient from the dataset may be greater than changing from $k$the value from k to $k-1$. Therefore, in aggregate numeric data, there may be multiple entries and values higher than 1 for a patient. So, the method for noise addition discussed above will not hold.**

If the target is the patient who visited the hospital five times, we get the values for the dataset with and without the target as k and (k-5). Thus, the ratio becomes:

$$\frac{Laplace(x|k, \frac{1}{\epsilon})}{Laplace(x|k-5, \frac{1}{\epsilon})} = e^{-\epsilon \cdot |x-k| - (-\epsilon \cdot |x-(k-5)|)}$$

In case 1, where $x < k-5$, we get:

$$\frac{Laplace(x|k, \frac{1}{\epsilon})}{Laplace(x|k-5, \frac{1}{\epsilon})} = e^{-5\epsilon}$$

In case 2, where $k-5 \leq x < k$, we get:

$$\frac{Laplace(x|k, \frac{1}{\epsilon})}{Laplace(x|k-5, \frac{1}{\epsilon})} = e^{2\epsilon \cdot (x-k) + 5\epsilon}$$

In case 3, where $x \geq k$, we get:

$$\frac{Laplace(x|k, \frac{1}{\epsilon})}{Laplace(x|k-5, \frac{1}{\epsilon})} = e^{5\epsilon}$$

Thus, we see that when $x < k-5$ or $x \geq k$, only 5ϵ-differential privacy is satisfied, while in the case $k-5 \leq x < k$, differential privacy is not satisfied. So, we use the initial value k as μ and $\frac{5}{\epsilon}$ as b. Using these values, we get the probability ratio as:

$$\frac{Laplace(x|k, \frac{5}{\epsilon})}{Laplace(x|k-5, \frac{5}{\epsilon})} = e^{\frac{-\epsilon \cdot |x-k| + \epsilon \cdot |x-(k-5)|}{5}}$$

We get three cases. In case 1, where $x < k-5$, we get:

$$\frac{Laplace(x|k, \frac{5}{\epsilon})}{Laplace(x|k-5, \frac{5}{\epsilon})} = e^{-\epsilon}$$

In case 2, where $k-5 \leq x < k$, we get:

$$\frac{Laplace(x|k, \frac{5}{\epsilon})}{Laplace(x|k-5, \frac{5}{\epsilon})} = e^{\frac{2\epsilon \cdot (x-k) + 5\epsilon}{5}}$$

In case 3, where $x \geq k$, we get:

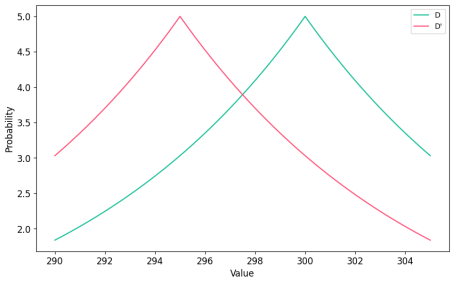
$$\frac{Laplace(x|k, \frac{5}{\epsilon})}{Laplace(x|k-5, \frac{5}{\epsilon})} = e^{\epsilon}$$

Therefore, we see that, if we substitute $\frac{5}{\epsilon}$ for b, we achieve differential privacy as long as we add noise from either the extreme left or the extreme right. This leads to our next result:

Result 2: When adding noise to data fields that contain aggregate numeric data, such as the number of patient visits to a hospital, we must add noise from a Laplace distribution with $b = \frac{n}{\epsilon}$, where n is the maximum value of the numerical variable.

The above result can be intuitively understood in figure 2, where $\epsilon = 0.5$ and $k = 300$ for D and $k = 295$ for D′:

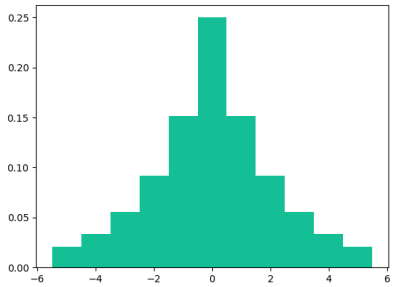
Fig. 2: **Noise Addition to Numeric Data**



In figure 2, if we were to take a point such as 302 again, we realise that the attacker cannot infer any data about the target as there is a possibility, albeit not equal, of the initial value being 295 or 300.

3) Integral Values for Noise: **As discussed in section II-A, a major natural constraint in healthcare date is that of integer values. For example, the aggregate number of cases of a disease must be integral. Adding noise to the value that makes it a decimal would immediately compromise the dataset, as an attacker can identify that it has been tampered with. Furthermore, fractional values for such integral variables may lead to problems in healthcare analytics on the data. As noise added to the data must be integral, we can either draw only integral values from a Laplace distribution or round random values drawn from a Laplace distribution to the nearest integer. Figure 3 shows the Laplace distribution when only integral values are drawn In the first case, we get the following change in the graph for :**

Fig. 3: **Drawing Only Integral Values**



In figure 3, the area of the graph using integral values, A, which is 1 in the usual Laplace distribution, equals:

$$A = \sum_{i=-\infty}^{\infty} \frac{1}{2b} \cdot e^{-\frac{|i-\mu|}{b}},$$

As the location parameter, μ only shifts the curve and doesn't affect area, we can use $\mu = 0$ to simplify as follows:

$$A = \frac{1}{2b} \cdot (e^{-\frac{|-\infty|}{b}} + ... + e^{-\frac{|-1|}{b}} + e^{-\frac{|0|}{b}} + e^{-\frac{|1|}{b}} + ... + e^{-\frac{|\infty|}{b}})$$

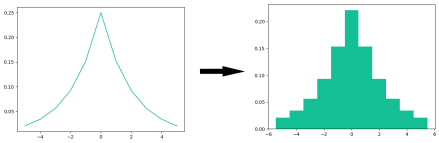
$$A = \frac{1}{b} \cdot (1 + e^{-\frac{1}{b}} + e^{-\frac{2}{b}} + e^{-\frac{3}{b}} + ... + e^{-\frac{\infty}{b}}) - \frac{1}{2b} = \frac{\epsilon}{2} \cdot \frac{e^{\epsilon}+1}{e^{\epsilon}-1}$$

The last step above is obtained using $b = \frac{1}{\epsilon}$. As the expression for the area is always greater than 1, it cannot be a viable Laplace distribution. This leads to the next important result:

Result 3: When noise is added to data fields containing integral values to satisfy differential privacy, taking only integer values from the Laplace distribution does not satisfy differential privacy.

So, we must use the second method of rounding the values taken from the Laplace distribution to the nearest integer. Using this method, the probability for each integral value, x, is the integral from $x - 0.5$ to $x + 0.5$. So, the Laplace distribution changes as shown in figure 4.

Fig. 4: **Rounding After Drawing Random Values**



Rounding randomly drawn values makes the above change in the graph, where each bar represents an integral of probabilities of its values. Therefore, the sum of the probabilities of all the bars does the same thing as applying an integral on the entire graph.

$$\sum_{i=-\infty}^{\infty} \int_{i-0.5}^{i+0.5} \frac{1}{2b} \cdot e^{-\frac{|x-\mu|}{b}} dx = \int_{-\infty}^{\infty} \frac{1}{2b} \cdot e^{-\frac{|x-\mu|}{b}} dx$$

Result 4: When noise is added to data fields containing integral values to satisfy differential privacy, rounding values drawn from a Laplace distribution provides differential privacy.

4) Noise Addition in Correlated Fields: **As discussed in section II-A, if independent distortions, or noise, are added to correlated fields in a dataset, the dataset itself may be compromised. Therefore, we must change our method for adding noise to correlated fields.**

To counteract this, we can add noise to only one of the fields, so that both fields will be distorted, but not independently. Furthermore, when one of the fields is fractional, the noise must be added to the other fields. This is because the fractional field is likely a composition of two fields. Therefore, adding noise to the fractional value creates disproportionate noise, or no noise at all, for the other fields. Thus, noise must be added to aggregate fields. Therefore, we get the following result:

Result 5: When adding noise to correlated fields, add it to only one of the fields so that the correlations hold. If one of the correlated fields is fractional, do not add the noise to that field as the noise may be disproportionately distributed.

III. Effect of Noise Addition on Analytics

In this section, we analyze the effect of noise addition, to satisfy differential privacy, on the accuracy of healthcare analytics. We focus on two techniques used for analysis of healthcare data: ordinary least squares regression and difference-in-difference estimation using panel data techniques. To perform these analyses, we assume that both x and y variables are confidential and, thus, noise must be added to both variables.

A. Theoretical Analysis To Assess The Impact Of Addition Of Noise On Quality Of Healthcare Analysis

Denote $L(\mu, b)$ as a function that finds a random value from the Laplace distribution with parameters μ and b. We use $L'(\mu, b)$ to denote a random draw that is different from $L(\mu, b)$. Adding noise to an ordinary least squares regression gives us the following equation

$$\left[y_i + L'\left(0, \frac{1}{\epsilon}\right)\right] = \beta_0 + \beta_1 \cdot \left[x_i + L\left(0, \frac{1}{\epsilon}\right)\right] + \nu_i,$$

where and i represents an iteration. We know the formula for β_1 is:

$$\beta_1 = \frac{cov(x, y)}{var(x)} \tag{1}$$

Therefore, the formula for β_1 after noise addition is:

$$\beta_1' = \frac{cov(x + L(0, \frac{1}{\epsilon}), y + L'(0, \frac{1}{\epsilon}))}{var(x + L(0, \frac{1}{\epsilon}))}$$

Now, using the fact that $L(0, \frac{1}{\epsilon})$ has mean 0 and it is independent of x so that its covariance with x is 0, the denominator is:

$$var[x + L(0, \frac{1}{\epsilon})] = E[(x - \overline{x})^2] = var(x) + var[L(0, \frac{1}{\epsilon})]$$

$$\therefore var[x + L(0, \frac{1}{\epsilon})] = var(x) + \frac{2}{\epsilon^2}$$

Now to simplify the numerator, we use the fact that $L(0, \frac{1}{\epsilon})$ and $L'(0, \frac{1}{\epsilon})$ have mean 0, they are independent of each other, and they are both independent of both x and y too. The independence leads to their pair-wise covariances being 0. So, the numerator is:

$$cov[x + L(0, \frac{1}{\epsilon}), y + L'(0, \frac{1}{\epsilon})] = cov(x, y)$$

Using the simplified numerator and denominator, we get

$$\therefore \beta_1' = \frac{cov(x, y)}{var(x) + \frac{2}{\epsilon^2}} \tag{2}$$

Using equations (1) and (2), we get:

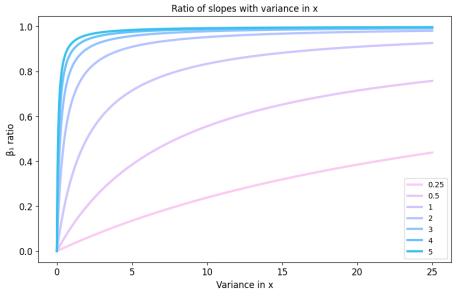
$$\frac{\beta_1}{\beta_1'} = 1 + \frac{2}{\epsilon^2 \cdot var(x)} \tag{3}$$

The ratio of slopes quantifies the accuracy of healthcare analyses following noise addition. So, the accuracy of the results from an analysis of noisy data is dependent on the values of ϵ and $var(x)$. Also, it is easy to observe that $\beta_1' < \beta_1$. Therefore, we get the following result:

Result 6: The slope parameter, which is used for health analytics, is lower in magnitude following the addition of noise when compared to the slope parameter in the original dataset.

If we plot $\frac{\beta_1'}{\beta_1}$ against variance in x over seven different values for epsilon, we get figure 5.

Fig. 5: **Change in Accuracy of Analytics Following Noise Addition**



In figure 5, we notice that, for a set value of epsilon, as the variance in x increases, $\frac{\beta_1'}{\beta_1}$ approaches 1. We also notice that, for a set variance in x, as the value for epsilon rises, $\frac{\beta_1'}{\beta_1}$ increases. As $\frac{\beta_1'}{\beta_1}$ approaches 1, β_1' approaches β_1. Therefore, the accuracy of healthcare analytics rises with rise in epsilon and with rise in variance in x.

Therefore we get the following two results:

Result 7A: The accuracy of analytics following noise addition increases with increases in both the value of ϵ and in the variance of the independent variable (x). Formally,

$$\frac{d\gamma}{d\epsilon} > 0, \ \frac{d\gamma}{d(var(x))} > 0.$$

Result 8A: The accuracy of analytics following noise addition increases disproportionately with increase in the value of ϵ when the variance of the independent variable (x) increases. Formally,

$$\frac{d^2\gamma}{d\epsilon \cdot d(var(x))} > 0.$$

B. Empirical Evidence

This subsection performs an experimental analysis of the effects of noise addition on healthcare analyses, specifically Ordinary Least squares (OLS) and Difference-in-difference (DID). We add noise to a real healthcare dataset containing data on vaccination and health outcomes across all the states in the United States. The data on health outcomes—which includes the total number of cases per million people, the total number of deaths per million people, and the percentage case fatality rate—is collected from Oxford University's COVID-19 Government Response Tracker. The data on vaccination is collected from ourworldindata.org. The time period of the data is from Jan-2021, when vaccination first began in the U.S. to Apr-2021, when we had collected the data. We add noise to this data using nine different values of epsilon: 0.25, 0.5, 1, 2, 3, 4, 5, 6, and 7. In both analyses, ordinary least squares and difference-in-difference techniques, we compare the results we obtain using the noisy dataset with the results obtained on the original dataset. We do this comparison for each value of epsilon to find the most accurate epsilon, the one with the least difference in the results vis-à-vis the original. We repeat this 100 times and aggregate the results to find the most accurate epsilon value. We also find the average difference from the original analysis over the 100 repetitions for each epsilon value to find the effects of noise addition.

The results of our analysis are shown below. Graphing a bar chart with the number of times each epsilon value was most accurate, we get figures 6 and 7. Figures 6 and 7 confirm our findings in the theoretical analysis, that an increase in the value of ϵ increases the accuracy of healthcare analytics.

Fig. 6: **Most Accurate ϵ for OLS Regression**

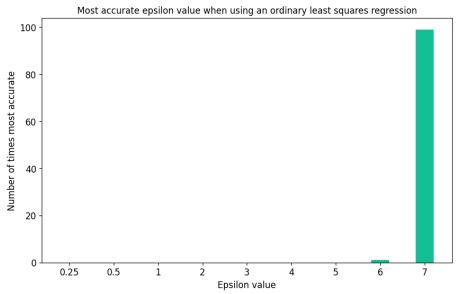
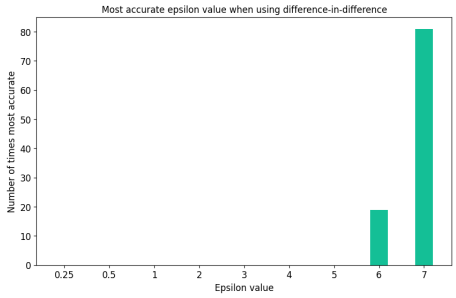


Fig. 7: **Most Accurate ϵ for Difference-in-difference Estimates**



Result 7B: As theoretically predicted in Result 7A, the empirical analysis using actual healthcare data confirms that the accuracy of analytics following noise addition increases with increase in the value of ϵ.

Graphing a bar chart with the average sum of squares error for each epsilon value in figures 8 and 9, we observe that the average difference across 100 iterations between the result obtained using the noisy dataset and the original dataset monotonically decreases with an increase in epsilon. This is consistent with what we found using the earlier figures 6 and 7 charts that displayed the most accurate epsilon.

Fig. 8: **Accuracy of Analytics using OLS Regression**

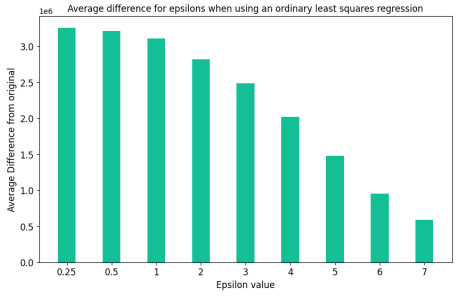
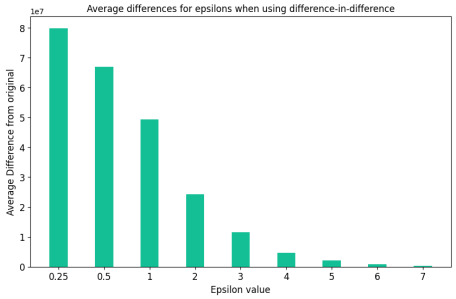


Fig. 9: **Accuracy of Analytics using Difference-in-difference Estimates**



The exact values for the average differences in millions are:

ϵ	OLS	DID	ϵ	OLS	DID
0.25	3.3	79.9	4	2.0	4.7
0.5	3.2	66.9	5	1.5	2.2
1	3.1	49.3	6	0.9	0.9
2	2.8	24.3	7	0.6	0.4
3	2.5	11.6			

If we look at the above values, we realize that as the value of epsilon rises, the average difference falls—by approximately 6 times for ordinary least squares and approximately 200 times for difference-in-difference. Thus, there is a disproportionately larger drop in the average differences for the difference-in-difference estimate when compared to the estimates obtained using the ordinary least squares. As seen in equation ?? for the coefficient estimate in a difference-in-difference analysis, the independent variable in a difference-in-difference analysis resembles the discrete equivalent of a second order derivative. Therefore, the variance of the independent variable in a difference-in-difference analysis is significantly greater than the variance of the independent variable in a ordinary least squares analysis. is consistent with that of the theoretical analysis. Result 7A predicts that as the variance of the independent variable increases, an increase in ϵ disproportionately increases the accuracy. Therefore, we find that the empirical evidence when comparing the change in accuracy with ϵ for the difference-in-difference analysis versus that for the ordinary least squares analysis is consistent with result 7A:

Result 8B: As theoretically predicted in Result 8A, the effect of an increase in the value of ϵ is disproportionately more in a difference-in-difference analysis than in an ordinary least squares regression.

IV. Conclusion

In a world that assigns increasing importance to privacy, especially in a confidential field such as healthcare, this paper analyses the potential applications of a new concept in privacy—differential privacy—to ensure data privacy in healthcare data. Differential privacy, a concept that enables provision of privacy to individuals in a dataset, can enable patients' privacy without compromising on the usability of the data. Our study first examines the techniques for adding noise to various types of healthcare data. We then examine the quality of analytics following addition of noise and confirm the predictions using an actual healthcare dataset.

Given the promise of differential privacy to preserving the privacy of individuals' data, a follow up to our study could be analyzing the various techniques to satisfy differential privacy at an individual level. Furthermore, as our study focuses on the use of Laplace distributions for adding noise to preserve differential privacy, subsequent work can explore other methods to satisfy

differential privacy and their applications to healthcare data. Subsequent work could also examine the methods to preserve differential privacy for other analytical techniques, such as artificial intelligence and machine learning.

REFERENCES

[1] L. Na, C. Yang, C.-C. Lo, F. Zhao, Y. Fukuoka, and A. Aswani, "Feasibility of reidentifying individuals in large national physical activity data sets from which protected health information has been removed with use of machine learning," *JAMA network open*, vol. 1, no. 8, pp. e186 040–e186 040, 2018.

[2] K. El Emam, E. Jonker, L. Arbuckle, and B. Malin, "A systematic review of re-identification attacks on health data," *PloS one*, vol. 6, no. 12, p. e28071, 2011.

[3] D. Barth-Jones, "The 're-identification' of governor william weld's medical information: a critical re-examination of health data identification risks and privacy protections, then and now," *Then and Now (July 2012)*, 2012.

[4] G. Simon, S. M. Shortreed, R. Y. Coley, and E. M. Iturralde, "Toolkit for assessing and mitigating risk of re-identification when sharing data derived from health records," 2020.

[5] C. Dwork, "Differential privacy," in *International Colloquium on Automata, Languages, and Programming*. Springer, 2006, pp. 1–12.

[6] S. A. Vinterbo, A. D. Sarwate, and A. A. Boxwala, "Protecting count queries in study design," *Journal of the American Medical Informatics Association*, vol. 19, no. 5, pp. 750–757, 2012.

[7] F. K. Dankar and K. El Emam, "Practicing differential privacy in health care: A review." *Trans. Data Priv.*, vol. 6, no. 1, pp. 35–67, 2013.

[8] J. Gardner, L. Xiong, Y. Xiao, J. Gao, A. R. Post, X. Jiang, and L. Ohno-Machado, "Share: system design and case studies for statistical health information release," *Journal of the American Medical Informatics Association*, vol. 20, no. 1, pp. 109–116, 2013.

[9] N. Mohammed, X. Jiang, R. Chen, B. C. Fung, and L. Ohno-Machado, "Privacy-preserving heterogeneous health data sharing," *Journal of the American Medical Informatics Association*, vol. 20, no. 3, pp. 462–469, 2013.

[10] C. Lin, Z. Song, H. Song, Y. Zhou, Y. Wang, and G. Wu, "Differential privacy preserving in big data analytics for connected health," *Journal of medical systems*, vol. 40, no. 4, pp. 1–9, 2016.

[11] M. Moussa and S. A. Demurjian, "Differential privacy approach for big data privacy in healthcare," in *Privacy and Security Policies in Big Data*. IGI Global, 2017, pp. 191–213.

[12] F. Prasser, F. Kohlmayer, H. Spengler, and K. A. Kuhn, "A scalable and pragmatic method for the safe sharing of high-quality health data," *IEEE journal of biomedical and health informatics*, vol. 22, no. 2, pp. 611–622, 2017.

[13] B. Yüksel, A. Küpçü, and Ö. Özkasap, "Research issues for privacy and security of electronic health services," *Future Generation Computer Systems*, vol. 68, pp. 1–13, 2017.

[14] F. Angeletti, I. Chatzigiannakis, and A. Vitaletti, "Towards an architecture to guarantee both data privacy and utility in the first phases of digital clinical trials," *Sensors*, vol. 18, no. 12, p. 4175, 2018.

[15] J. W. Kim, B. Jang, and H. Yoo, "Privacy-preserving aggregation of personal health data streams," *PloS one*, vol. 13, no. 11, p. e0207639, 2018.

[16] H.-A. Park, "The modeling of privacy preserving and statistically analysable database (ppsadb) system," *International Journal of Advanced Computer Research*, vol. 8, no. 38, pp. 229–239, 2018.

[17] J. L. Raisaro, J. R. Troncoso-Pastoriza, M. Misbach, J. S. Sousa, S. Pradervand, E. Missiaglia, O. Michielin, B. Ford, and J.-P. Hubaux, "M ed c o: Enabling secure and privacy-preserving exploration of distributed clinical and genomic data," *IEEE/ACM transactions on computational biology and bioinformatics*, vol. 16, no. 4, pp. 1328–1341, 2018.

[18] S. Banerjee, R. Benlamri, and S. Bouzefrane, "Optimization of ontology-based clinical pathways and incorporating differential privacy in the healthcare system," *Security designs for the cloud, iot, and social networking*, pp. 191–205, 2019.

[19] H. Cho, S. Simmons, R. Kim, and B. Berger, "Privacy-preserving biomedical database queries with optimal privacy-utility trade-offs," *Cell systems*, vol. 10, no. 5, pp. 408–416, 2020.

[20] D. Froelicher, M. Misbach, J. R. Troncoso-Pastoriza, J. L. Raisaro, and J.-P. Hubaux, "Medco2: privacy-preserving cohort exploration and analysis," in *Digital Personalized Health and Medicine*. IOS Press, 2020, pp. 317–321.

[21] D. R. Harris, "Leveraging differential privacy in geospatial analyses of standardized healthcare data," in *2020 IEEE International Conference on Big Data (Big Data)*. IEEE, 2020, pp. 3119–3122.

[22] R. Bild, K. A. Kuhn, and F. Prasser, "Better safe than sorry–implementing reliable health data anonymization," in *Digital Personalized Health and Medicine*. IOS Press, 2020, pp. 68–72.

[23] A. Dyda, M. Purcell, S. Curtis, E. Field, P. Pillai, K. Ricardo, H. Weng, J. C. Moore, M. Hewett, G. Williams *et al.*, "Differential privacy for public health data: An innovative tool to optimize information sharing while protecting data confidentiality," *Patterns*, vol. 2, no. 12, p. 100366, 2021.

[24] J. Ficek, W. Wang, H. Chen, G. Dagne, and E. Daley, "Differential privacy in health research: A scoping review," *Journal of the American Medical Informatics Association*, vol. 28, no. 10, pp. 2269–2276, 2021.

[25] A. Hägermalm and S. Slavnic, "Differential privacy: an extensive evaluation of open-source tools for ehealth applications," 2021.

[26] A. W, Huang, A. Kandula, and X. Wang, "A differential-privacy-based blockchain architecture to secure and store electronic health records," in *2021 The 3rd International Conference on Blockchain Technology*, 2021, pp. 189–194.

[27] K. M. Chong and A. Malip, "Bridging unlinkability and data utility: Privacy preserving data publication schemes for healthcare informatics," *Computer Communications*, 2022.