

Article

Child Health Dataset Publishing and Mining Based on Differential Privacy Preservation

Wenyu Li ¹, Siqi Wang ¹, Hongwei Wang ²  and Yunlong Lu ^{1,*} 

¹ School of Mathematics and Statistics, Beihua University, Jilin 132013, China; liwenyu82@163.com (W.L.); 15568359926@163.com (S.W.)

² Departments of Mathematics and Physics, Texas A&M International University, Laredo, TX 78045, USA; hongwei.wang@tamiu.edu

* Correspondence: luyunlong78@gmail.com

Abstract: With the emergence and development of application requirements such as data analysis and publishing, it is particularly important to use differential privacy protection technology to provide more reliable, secure, and compliant datasets for research in the field of children's health. This paper focuses on the differential privacy protection of the ultrasound examination health dataset of adolescents in southern Texas from three aspects: differential privacy protection with output perturbation on basic statistics, publication of differential privacy marginal histogram and synthesized data, and a machine learning differential privacy learning algorithm. Firstly, differential privacy protection results with output perturbation show that Laplace and Gaussian mechanisms for numerical data, as well as the exponential mechanism for non-numerical data, can achieve the goal of protecting privacy. The exponential mechanism provides higher privacy protection. Secondly, a differential privacy marginal histogram with four attributes can be obtained with an appropriate privacy budget that approximates the marginal histogram of the original data. In order to publish synthetic data, we construct a synthetic query to obtain the corresponding differential privacy histogram for two attributes. Further, a synthetic dataset can be constructed by following the data distribution of the original dataset and the quality of the synthetic data publication can also be evaluated by the mean square error and error rate. Finally, consider a differential privacy logistic regression model under machine learning to predict whether children have fatty liver in binary classification tasks. The experimental results show that the model combined with quadratic perturbation has better accuracy and privacy protection. This paper can provide differential privacy protection models under different demands, which provides important data release and analysis options for data managers and research organizations, in addition to enriching the research on child health data releasing and mining.

Keywords: children health; synthetic data release; differential privacy; marginal histograms; logistic regression

MSC: 68P27



Citation: Li, W.; Wang, S.; Wang, H.; Lu, Y. Child Health Dataset Publishing and Mining Based on Differential Privacy Preservation. *Mathematics* **2024**, *12*, 2487. <https://doi.org/10.3390/math12162487>

Academic Editor: Cheng-Chi Lee

Received: 4 June 2024

Revised: 2 August 2024

Accepted: 6 August 2024

Published: 12 August 2024



Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

The rapid collection of digital information has led to a proliferation of datasets over the past two decades, and the rate of collection has become even more rapid in recent years. Data managers are responsible for releasing data for further analysis. Figure 1 depicts the two phases of data collection and data release/analysis. In the data collection phase, individuals submit personal information to form datasets that contain sensitive information, and data managers have full control over the datasets. In the data release/analysis phase, the aim is to share the dataset or query results with the public, providing data models and specific machine learning or data mining algorithms [1]. Data release and analysis

bring social benefits such as providing quality services, publishing official statistics, and supporting data mining or machine learning tasks. However, the importance of data privacy protection comes to the fore. If data managers are not trustworthy, personal information may be leaked directly during the data collection phase. Even if the data manager is trustworthy and adopts anonymization techniques, when releasing aggregated information to the public, personal information may still be at risk of leakage due to the untrustworthy nature of the general public. Therefore, privacy protection is a vital issue.

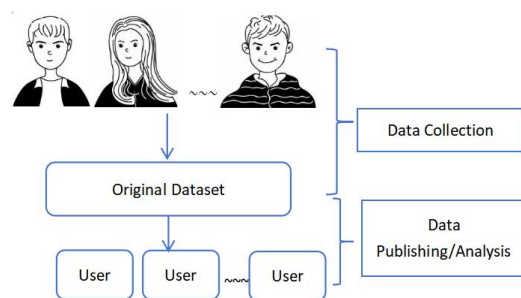


Figure 1. Data collection and data dissemination/analysis.

Differential privacy has emerged as a response to increasing data privacy issues. It protects individuals' privacy by introducing a certain amount of noise or perturbation into data processing, making it difficult for attackers to retrieve sensitive information. At the same time, it is committed to maintaining the utility and usability of data. Differential privacy was first proposed by Dwork et al. in 2006 [2], who laid the theoretical foundation for it and demonstrated its algorithms and application to specific tasks [3,4]. It provides various solutions for data privacy protection, such as querying, histogram publishing [5], synthetic data publishing [6], and learning algorithms [7].

This article aims to explore differential privacy protection in a children's health dataset, with a focus on three main issues: (a) differential privacy protection based with the three mechanisms on basic statistics, (b) publishing based on marginal histograms and composite datasets, and (c) differential privacy protection in machine learning algorithms based on adding noise to the loss function. Finally, a conclusion is included.

2. Preliminary Knowledge

2.1. Definitions Related to Differential Privacy

Assuming a finite data space X , let r represent a record with d attributes, and the dataset is an unordered set sampled from the finite data space. If there is only one record r that is different between two datasets D and D' , then D and D' are called adjacent datasets. Query f is a function that maps the dataset D to the range set R .

Differential privacy: M provides (ϵ, δ) -differential privacy for each set of outputs Ω in D and any neighboring dataset of D' if the randomization mechanism M satisfies the following condition: $\Pr[M(D) \in \Omega] \leq e^\epsilon \times \Pr[M(D') \in \Omega] + \delta$.

Sensitivity: For a query $f: D \rightarrow R$, and the neighboring datasets D and D' , the ℓ_1 -sensitivity Δf of f is defined as $\Delta f = \max_{D, D'} \|f(D) - f(D')\|_1$.

2.2. Three Differential Privacy Protection Mechanisms

Differential privacy-preserving mechanisms typically use Laplace and Gaussian mechanisms for numerical queries and exponential mechanisms for non-numerical queries.

Consider

$$M(D) = f(D) + Y. \quad (1)$$

The Laplace mechanism M provides ϵ -differential privacy ($\delta = 0$) if $Y \sim \text{Lap}(\frac{\Delta f}{\epsilon})$, where $\text{Lap}(x) = \frac{1}{2b} \exp(-\frac{|x|}{b})$; the Gaussian mechanism M satisfies (ϵ, δ) -differential privacy

if $Y \sim N(0, \sigma^2)$ follows normal distribution, and for any $\delta \in (0, 1)$, $\sigma > \frac{\sqrt{2 \ln \frac{2}{\delta}} \Delta f}{\epsilon}$. The exponential mechanism M satisfies ϵ -differential privacy when the following conditions are satisfied:

$$M(D) = \text{return } \varphi \propto \exp\left(\frac{\epsilon q(D, \varphi)}{2\Delta q}\right), \quad (2)$$

where $q(D, \varphi)$ is the score function of dataset D , which measures the quality of output $\varphi \in \Omega$, the sign $\propto$ indicates proportionality, and Δq denotes the sensitivity of φ .

3. Differential Privacy Protection for the Children's Health Dataset

In this study, the children's health dataset adopts the data processing method described in reference [8]. The hardware environment used in the experiment is 7th Gen Intel (R) Core (TM) i5 2.60 GHz, 8 GB of memory, implemented using Python 3.9 programming language, and the operating system platform is Windows 10.

3.1. Differential Privacy Protection for Basic Statistics

Next, three differential privacy noise mechanisms will be used to analyze and evaluate the utility issues of querying children with certain diseases. More specifically, for non-numerical exponential mechanisms, we will construct a rating function suitable for the dataset to achieve privacy protection.

According to the medical standard, people suffer from obesity when their BMI $> 28 \text{ kg/m}^2$ and kidney disease when the glomerular filtration rate is outside of 90–120 mL/min. People suffering from fatty liver disease and hepatomegaly are represented in the Excel data sheet by 1 and 0, respectively.

Privacy protection is based on Formulas (1) and (2) with different ϵ and $\Delta f = 1$, the specific number of patients are shown in Tables 1 and 2, and the probability of having diseases is listed in Table 3.

Table 1. Laplace mechanism output.

Disease	# of People	$\epsilon = 0.01$	$\epsilon = 0.1$	$\epsilon = 1$
Fatty liver	366	221	343	365
Obesity	466	424	474	468
Nephropathy	804	769	800	806
Hepatomegaly	245	214	240	244

Table 2. Gaussian mechanism output.

Disease	# of People	$\epsilon = 0.01$	$\epsilon = 0.1$	$\epsilon = 1$
Fatty liver	366	761	196	368
Obesity	466	459	472	467
Nephropathy	804	720	738	806
Hepatomegaly	245	414	291	254

Table 3. Exponential mechanism output.

Disease	# of People	$\epsilon = 0.01$	$\epsilon = 0.1$	$\epsilon = 1$
Fatty liver	366	0.25	1.02×10^{-8}	1.23×10^{-80}
Obesity	466	0.25	1.24×10^{-6}	8.63×10^{-60}
Nephropathy	804	0.25	1.0×10^0	1.0×10^0
Hepatomegaly	245	0.25	2.41×10^{-11}	6.53×10^{-107}

From Tables 1–3, it can be concluded that as the privacy budget ϵ decreases, the degree of privacy protection improves, but the data utility decreases. When ϵ is larger, the data utility is higher, and the degree of privacy protection is lower. From Tables 1 and 2, the

Gaussian mechanism for privacy protection has better privacy results compared to the Laplace mechanism, but the data availability of query results is poor.

For non-numerical queries using the exponential mechanism, the query function is “What is the most common disease in this region?” The scoring function q is taken as the number of people for each disease and the output probability is calculated according to what is shown in Table 3. When $\epsilon = 0.01$ or $\epsilon = 1$, the query results cannot properly balance privacy protection and data utility. When $\epsilon = 0.1$, the probability of selecting Nephropathy is the highest, and the probability of selecting Hepatomegaly is the lowest. This indicates that the mechanism can balance privacy and data utility under the some precision of privacy protection.

3.2. Data Publication

3.2.1. Marginal Histogram Release

Differential privacy marginal histograms are a differential privacy-preserving method for publishing and sharing data. This method allows users to access statistical information about the distribution of data while protecting the privacy of individual data. A marginal histogram is a graphical representation of the distribution of data that shows the frequency or probability of the data over a range of values. The goal of publishing differential privacy marginal histograms is to provide a rough understanding of data distribution without exposing specific individual data [9], as described by Figure 2 below.

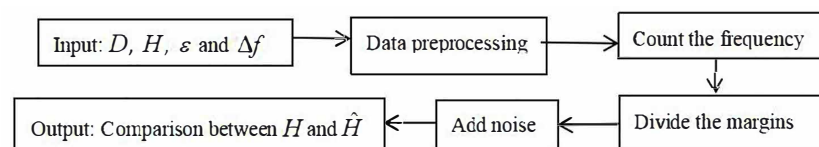


Figure 2. Differential privacy margin histogram publishing process. where D is the original dataset, $H = \{H_1, \dots, H_i, \dots, H_n\}$ is the original marginal histogram, where H_i denotes the number of frequencies in each bracket of the histogram, and $\hat{H}$ is the marginal histogram after adding noise.

The specific steps are as follows:

Step 1. Select four features: AGE, BMI, LIVER SIZE, and GFR.

Step 2. Divide the data into 30 brackets, then calculate and draw the marginal histogram of each feature.

Step 3. Use the Laplace mechanism to add noise to the frequency of each feature’s marginal histogram, and use an average allocation strategy for privacy budgeting.

Step 4. Draw the original marginal histogram and differential privacy protection histogram under different privacy budgets for comparison.

The histograms in Figures 3 and 4 are obtained according to the above steps.

Similar to the analysis in the previous section, Figures 3 and 4 indicate that when $\epsilon = 0.1$, the added noise to some extent disturbs the data, and the distribution of data with differential privacy remains basically consistent with the original data distribution, effectively protecting individual privacy.

To evaluate the quality of the publication, the mean squared error (MSE) formula [10],

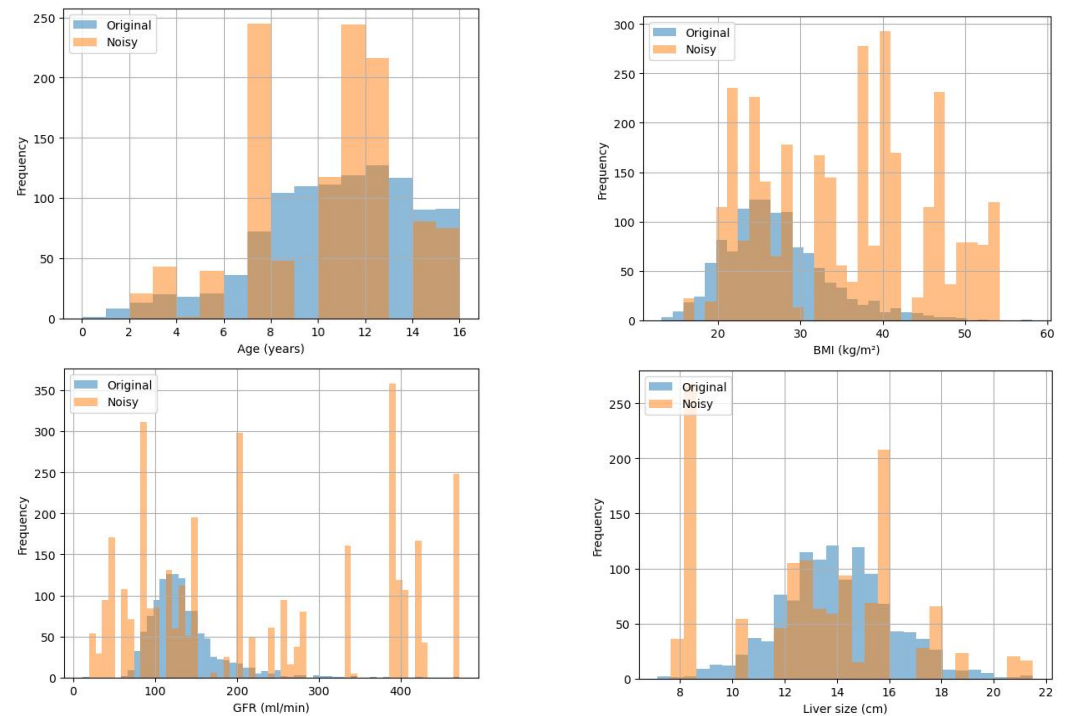
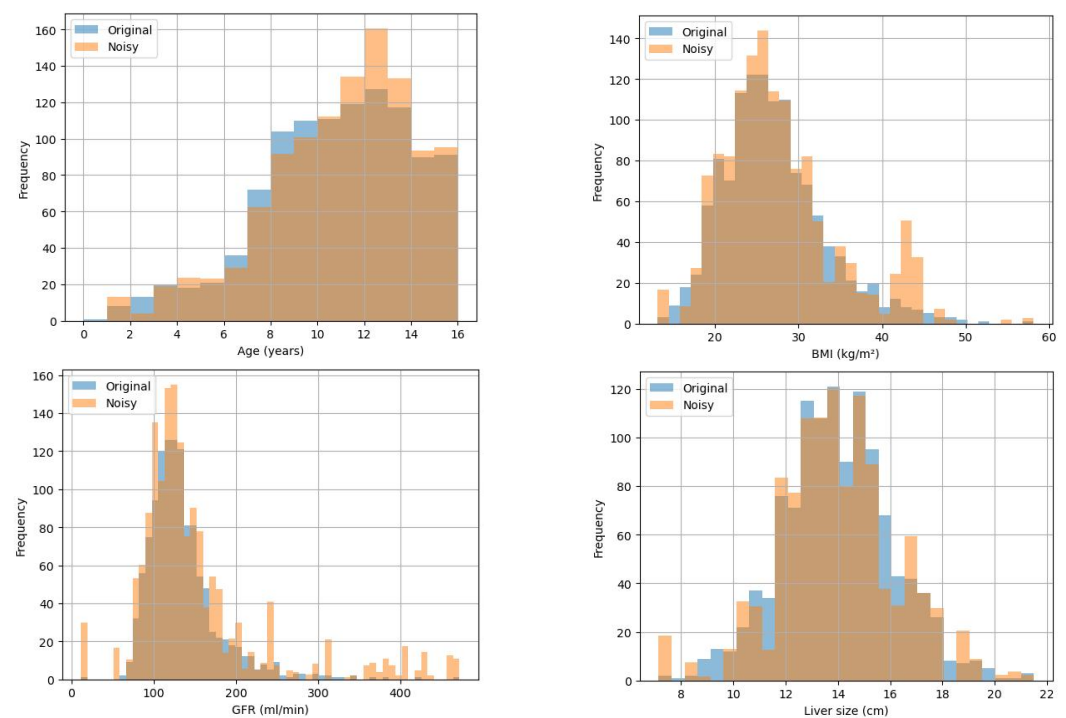
$$MSE = \frac{1}{n} \sum (H(x) - \hat{H}(x))^2, \quad (3)$$

is used to measure the difference between the differential privacy histogram and the original histogram, with the results shown in Table 4 below.

From Table 4, the AGE attribute has the lowest MSE, while the GFR attribute has a relatively low MSE. This indicates that using the same privacy budget for these attributes can better protect the privacy of AGE attributes and provide lower prediction errors. However, the MSEs of BMI and LIVER SIZE attributes are slightly higher under the same privacy budget, requiring a delicate balance between privacy protection and data accuracy.

Table 4. Mean squared error (MSE).

	$\epsilon = 0.5$	$\epsilon = 1$	$\epsilon = 1.5$	$\epsilon = 2$
Age (years)	4.4306	1.2782	0.8041	0.3199
BMI (kg/m ²)	7.0647	1.9203	0.8292	0.3623
Liver size (cm)	9.5685	1.9333	0.9142	0.3752
GFR (mL/min)	5.3761	1.6877	0.8873	0.3264

**Figure 3.** Marginal histograms of 4 attributes when $\epsilon = 0.01$.**Figure 4.** Marginal histograms of 4 attributes when $\epsilon = 0.1$.

3.2.2. Composite Dataset Publishing

Synthetic data publishing predicts potential distribution structures from raw data through generative models, generating data with similar statistical characteristics. Compared to the possibility of data distortion caused by differential privacy histogram publishing, synthetic data publishing can more effectively protect the privacy of data subjects and handle the relationships between different attributes to improve the utility of data. The key to its release is to construct a reasonable range of synthetic data queries. We adopt the approach of constructing a composite dataset that follows the data distribution of the original dataset to protect privacy [11,12]. The specific steps are as follows:

Step 1. Design a synthetic representation of answer range queries for a column of raw data showing the data distribution with a histogram.

Step 2. The count results in the histogram being used as a synthetic representation of the raw data.

Step 3. Laplace noise (2) is added individually to each count value in the histogram.

Step 4. The synthetic representation is used as a probability distribution function for sampling to generate the synthetic dataset.

Step 5. A new sample is generated based on the probability values.

Step 6. The generated distribution is compared with the distribution of the original dataset.

The results of these specific inquiries are as follows.

Firstly, range questioning under BMI and systolic blood pressure (SBP) attributes is designed to obtain the number of people suffering from obesity, with a BMI of 28–60 (kg/m^2), and hypertension, with an SBP of 140–200 (mmHg).

Defining the count value of each BMI between 28 and 60 as a histogram query and applying the range query to calculate the number of people per BMI yields Figure 5. Similarly, defining the count value of each SBP between 140 and 200 as a histogram query and applying the range query to calculate the number of people per SBP yields Figure 5.

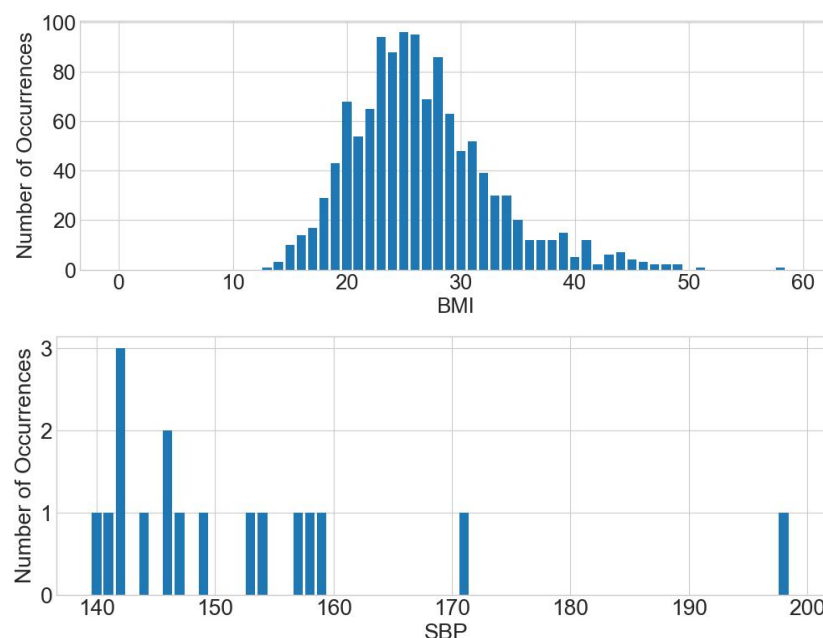


Figure 5. People with BMI 28–60 (kg/m^2) and SBPs 140–200 (mmHg).

Secondly, to answer the query range, the results of the BMI and SBP counts of all people who fall within the range are summed up. The results of the experiment show that there are 466 and 17 people who meet the above ranges for BMI and SBP, respectively.

Thirdly, according to the parallel combination property of differential privacy and Equation (1), the synthesized representation satisfies ϵ -differential privacy, and the results after noise addition are 465.990 and 16.524, respectively.

Fourthly, the composite representation as a probability distribution function that can be used to estimate the potential distribution of the original data is considered and then sampled based on this probability distribution to obtain a composite dataset. Comparing Figures 5 and 6, the probability distribution of the BMI and SBP estimates is similar to the probability distribution function of the original dataset.

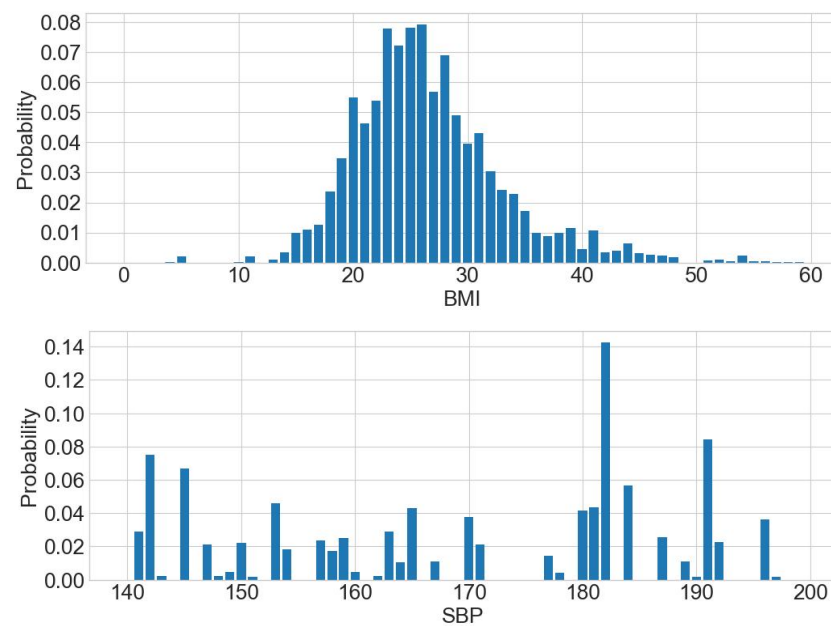


Figure 6. Probability distribution of BMI and SBP.

The composite dataset is obtained, as shown in Table 5.

Table 5. The composite dataset on BMI and SBP.

Serial	BMI (kg/m ²)	SBP (mmHg)
0	25	166
1	33	171
2	21	184
3	37	147
⋮	⋮	⋮
1212	24	142
1213	33	158
1214	25	175

The original dataset is then replaced with the generated synthetic dataset and queries are responded to based on it. The histograms of BMI and SBP for the synthetic dataset reveal graphs that are the same shape as the original dataset, as shown in Figure 7.

Next, to evaluate the quality of the synthesized data, the MSE Formula (3) is used, and the results are shown in Table 6.

Table 6 clearly shows that the synthetic data are more usable while protecting data privacy. Thus, the MSE can guide the choice of the synthetic data generation mechanism and control the cost of privacy protection.

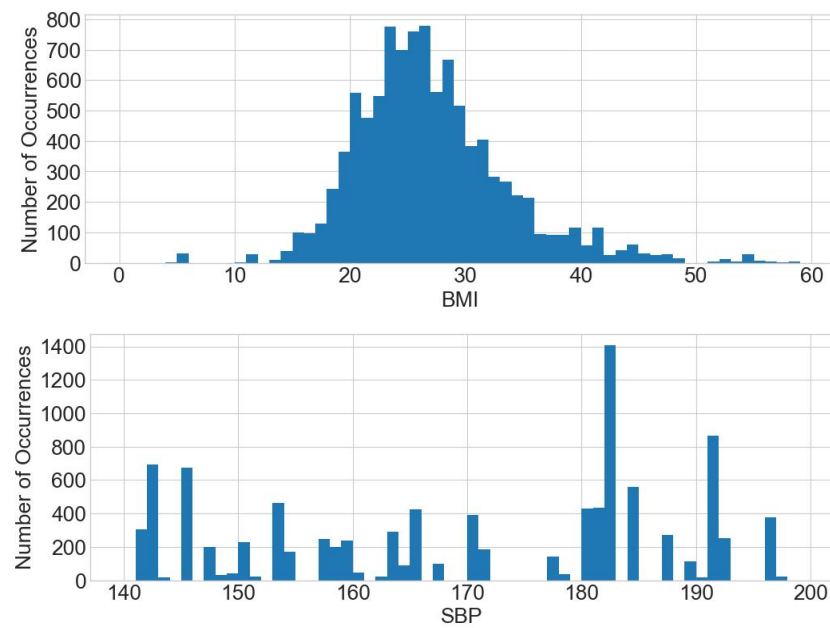


Figure 7. Histograms of BMI and SBP for the synthetic dataset.

Table 6. MSE of BMI and SBP ($\epsilon = 1$).

Feature	MSE
BMI	2.03362×10^{-3}
SBP	7.5318×10^{-4}

Further, to evaluate query accuracy, two error rates are defined, as follows:

$$e_l = \frac{|(\text{noisy query result}) - (\text{true query result})|}{\text{true query result}}, \quad (4)$$

where notation e_l denotes the Laplace mechanism error rate.

$$e_s = \frac{|s - o|}{o}, \quad (5)$$

where notation e_s denotes the synthetic representation error rate, s denotes the statistical characteristics of the synthesized data, and o denotes the statistical characteristics of the original data.

In Table 7, the smaller error rate indicates more accurate queries and the Laplace mechanism has mostly smaller errors. However, for queries with a larger range, the query cost increases and the error significantly increases. In practical applications, Laplace queries are typically used for small samples and budget-constrained queries, while composite data queries are used for large samples and large-scale queries.

Table 7. Error rates of queries.

Query Attribute	Query Range	e_s	e_l
BMI	30–31 (kg/m ²)	0.04526	0.02340
	30–71 (kg/m ²)	0.10352	0.09374
SBP	140–145 (mmHg)	0.76525	0.56056
	140–200 (mmHg)	1.56074	1.61417

3.2.3. Differential Privacy Protection for Machine-Learning Algorithms

This section explores the application of the logistic regression model with differential privacy in children's health datasets. In 2011, Chaudhuri et al. [7] established a differential privacy (CDP for short) theory based on target perturbation by adding linear noise to the loss function and applied it to the logistic regression model. Inspired by their study, Mi [13] added quadratic noise to the loss function to obtain a differential privacy protection model (MDP for short), then applied it to the classification task of children with fatty livers based on the BMI attributes in the dataset.

The dataset is divided into a training set and testing set, which are used to train the model and evaluate the testing accuracy, respectively. The accuracy formula is as follows [14]:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}, \quad (6)$$

where TP , TN , FP , and FN represent the number of patient true positive examples, true negative examples, false positive examples, and false negative examples, respectively. The experimental results of CDP and MDP are shown in Figures 8 and 9, where the values of ϵ are taken in the range of $[10^{-5}, 10^{10}]$ for 100 different points. The benchmark accuracy of the logistic regression model without privacy protection is about 0.7577.

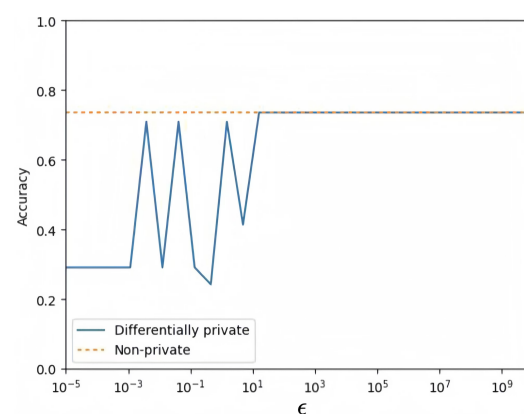


Figure 8. CDP logistic regression model.

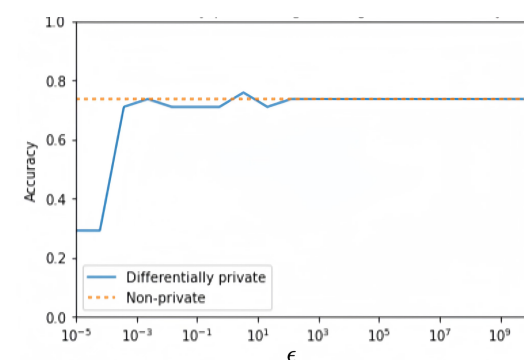


Figure 9. MDP logistic regression model.

By comparing the accuracy curves of Figures 8 and 9, it is found that the MDP model performs more stably and has higher accuracy, as shown in Figure 10. This indicates that under the same privacy budget, the MDP model has better privacy protection and higher model accuracy. Meanwhile, with the increase in the privacy budget, both models converge to the benchmark accuracy of privacy-free models.

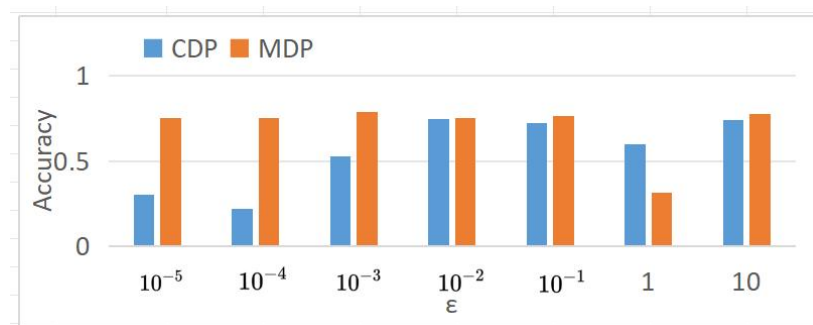


Figure 10. Comparison of prediction accuracy under different ϵ .

4. Conclusions

This paper implements the privacy protection of children's health datasets; provides querying for basic statistics of a dataset, histogram release, synthetic data release, and a differential privacy machine learning algorithm, which can effectively balance the utility of children's health data and privacy protection issues; and explore the advantages and disadvantages of different privacy protection mechanisms, aiming to provide a usage basis and reference for personnel with different levels of needs. However, for the problem of querying and releasing high-dimensional large samples, it is necessary to combine the correlation between attributes to construct the privacy mechanism, such as DPCopula technology.

Author Contributions: Conceptualization, Y.L.; methodology, W.L.; resources, H.W.; writing—original draft, S.W.; writing—review and editing, H.W. All authors have read and agreed to the published version of the manuscript.

Funding: This work is supported in part by the doctoral research project establishment and start-up fund of Beihua University, Jilin Provincial Department of Education, science and technology research projects “jjkh20220037kj” and “jjkh20220038kj”, Jilin Provincial Natural Science Foundation “YDZJ202201ZYTS320”.

Data Availability Statement: The original contributions presented in the study are included in the article, further inquiries can be directed to the corresponding author.

Acknowledgments: The authors show their appreciation to Francisco J. Cervantes, a Pediatrician in South Texas, for providing the data.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Zhu, T.; Li, G.; Zhou, W.; Philip, S.Y. *Differential Privacy and Applications*; Springer International Publishing: Cham, Switzerland, 2017.
2. Dwork, C. Differential privacy. In *International Colloquium on Automata, Languages, and Programming*; Springer: Berlin/Heidelberg, Germany, 2006; pp. 1–12.
3. Dwork, C. Differential privacy: A survey of results. In *Proceedings of the International Conference on Theory and Applications of Models of Computation*, Xi'an, China, 25–29 April 2008; Springer: Berlin/Heidelberg, Germany, 2008.
4. Dwork, C.; Roth, A. The algorithmic foundations of differential privacy. *Found. Trends Theor. Comput. Sci.* **2014**, *9*, 211–407. [\[CrossRef\]](#)
5. Xu, J.; Zhang, Z.; Xiao, X.; Yang, Y.; Yu, G. Differentially private histogram publication. *Vldb J.* **2013**, *22*, 797–822. [\[CrossRef\]](#)
6. Ghatak, D.; Sakurai, K. A survey on privacy preserving synthetic data generation and a discussion on a privacy-utility trade-off problem. In *Proceedings of the International Conference on Science of Cyber Security*, Matsue, Japan, 10–12 August 2022; Springer Nature: Singapore, 2022.
7. Chaudhuri, K.; Monteleoni, C.; Sarwate, A. Differentially private empirical risk minimization. *J. Mach. Learn. Res.* **2011**, *12*, 1069–1109. [\[PubMed\]](#)
8. Lu, Y.; Li, W.; Gong, X.; Mi, J.; Wang, H.; Quintana, F.G. Prevalence of fatty liver among children under multiple machine learning models. *South. Med. J.* **2022**, *115*, 622–627. [\[CrossRef\]](#) [\[PubMed\]](#)
9. Li, H.; Xiong, L.; Jiang, X. Differentially private histogram and synthetic data publication. In *Medical Data Privacy Handbook*; Springer: Cham, Switzerland, 2015; pp. 35–58.

10. Murphy, A.H. Skill scores based on the mean square error and their relationships to the correlation coefficient. *Mon. Weather. Rev.* **1988**, *116*, 2417–2424. [[CrossRef](#)]
11. Xin, B.; Geng, Y.; Hu, T.; Chen, S.; Yang, W.; Wang, S.; Huang, L. Federated synthetic data generation with differential privacy. *Neurocomputing* **2022**, *468*, 1–10. [[CrossRef](#)]
12. Near, J.; Abuah, C. Programming Differential Privacy. 2021. Available online: <https://uvm-plaid.github.io/programming-dp/> (accessed on 10 May 2024).
13. Mi, J. Research and Application of Empirical Risk Minimization Linear Model Based on Differential Privacy. Master's Thesis, Beihua University, Jilin, China, 2023.
14. Uddin, S.; Haque, I.; Lu, H.; Moni, M.; Gide, E. Comparative performance analysis of K-nearest neighbour algorithm and its different variants for disease prediction. *Sci. Rep.* **2022**, *12*, 6256. [[CrossRef](#)]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.