BMC Medical Informatics and Decision Making

RESEARCH

Open Access

Enhancing privacy protection of physical examination data through synthetic algorithms based on differential privacy

Weili Zhang[1], Ran Liu[2], Xinyi Zhu[3], Xiaojin Yu[1*] and Depeng Jiang[4*]

Abstract

Background Health physical examinations play a crucial role in early detection of cancer and chronic disease. However, privacy concerns limit the utilization of this kind of data for health interventions and research. Synthetic data methods based on differential privacy are increasingly used to create complete datasets that protect privacy while enabling data analysis and result interpretation. Hence, the use of synthetic algorithms based on differential privacy for privacy protection of physical examination data is a promising research direction.

Methods Three synthetic algorithms, PrivBayes, PeGS, and DP-Gibbs were used to generate complete synthetic datasets that adhere to differential privacy standards using physical examination data composed of categorical data, which compared with the existing algorithm Private-PGM.

Results Compared with the existing algorithm, DP-Gibbs can provide privacy preserving capacity of 4.686 ($\varepsilon = 0.5$), while the existing algorithm only with 2.012. In addition, DP-Gibbs provides 0.620 of precision, 0.539 of F1-score, 0.342 of Kappa Coefficient, and 0.765 of AUC-score. The corresponding statistical results of existing algorithm are 0.520, 0.321, 0.188 and 0.695.

Conclusions The main contributions of this study are the exploration of combination models incorporating different noise forms and Bayesian synthetic algorithms, alongside a comparative analysis against existing algorithms. This study explored the balance between privacy protection and data utility under different levels of privacy protection, and DP-Gibbs offers more stable technical support for de-identifying physical examination data prior to sharing and analysis, which realized the mining and application of a wider range of medical data under the requirements of privacy protection. By leveraging this effective privacy protection technique, clinical researchers can extract valuable insights on diseases and population health from the physical examination data without the risk of leaking private information.

Keywords Differential privacy, Synthetic dataset, Physical examination data, Bayesian network, Gibbs sampling

*Correspondence:
Xiaojin Yu
xiaojinyu@seu.edu.cn
Depeng Jiang
depeng.jiang@umanitoba.ca
[1]Department of Epidemiology and Health Statistics, School of Public Health, Southeast University, Nanjing 210009, China
[2]Department of Occupational and Environmental Health, School of Public Health, Southeast University, Nanjing 210009, China
[3]Zhongda Hospital Affiliated to Southeast University, Gulou District, Nanjing 210009, China
[4]Department of Community Health Sciences, University of Manitoba, Winnipeg, MB R3T 2N2, Canada

Background

With the rapid development of society's economy, there has been an increase in public' health literacy and awareness of disease prevention. Consequently, health physical examinations have become more prevalent, leading to a great number of physical examination data. These examinations are crucial for detecting early signs of chronic diseases like cancer, coronary heart disease (CHD) in seemingly healthy populations, and they provide essential guidance on further health interventions [1, 2]. However, the release of raw data by data collection organizations, such as hospitals, containing individuals' private and sensitive information pose a risk of privacy breaches during data mining. This situation can potentially lead to ethical and legal concerns for researchers.

Privacy-preserving algorithms are essential for handling data containing sensitive information. Techniques like *k*-anonymity, which involve masking and generalization, along with its derivatives *l*-diversity and *T*-closeness, are widely used for privacy protection [3–7]. These techniques protect privacy by constraining the link between quasi-identifiers and sensitive information. Based on these anonymization models, some studies have also proposed privacy protection ideas combined with local suppression techniques and local generalizations to avoid privacy leakage of sensitive data with missing values [8]. Considering that generalization leads to the reduction of data information, researchers have proposed the concept of anatomy, dividing the data set into a quasi-identifier table and a sensitive table to safeguard privacy and data correlation [9, 10]. Aggregate query frameworks constitute a valuable privacy protection idea as well. Specifically, the query condition is determined by quasi-identifier attributes, and the appropriate aggregate function is used to query each query result from the specific sensitive attribute. Datasets that meet these limitations can be granted privacy protection. Some researchers have proposed a privacy protection model for rating datasets, so called as $(l^{p1}, \ldots, l^{pn})$-privacy. The model provides datasets through the aggregate query framework, which guarantees that all possible query results have particular different sensitive values [11, 12]. Additionally, the concept of data shuffling is also used for privacy protection. Such as based on aggregate query frameworks, the values in each attribute of each partition are shuffled [13, 14]. However, these privacy-preserving algorithms rely on certain assumptions about the attacker's background knowledge and the attack model, which may not align with the real situation, leading to potential privacy breaches. In 2006, Dwork et al. introduced differential privacy as a mathematical framework to quantify the level of privacy protection. This model involves adding specific noise to query results when performing operations like summing and counting on datasets containing sensitive information. Importantly, differential privacy ensures that adding or removing any record from the dataset will not significantly impact the query results, thus avoiding the leakage of individual privacy information [15, 16].

Differential privacy algorithms are extensively employed within the context of data in cloud environment. For instance, prior to data sharing, *k*-anonymity can be utilized for data partition, followed by the injection of noise into sensitive data, thereby enabling synthetic data for classification prediction [17]. Furthermore, differential privacy algorithms can be integrated with classification models. By optimizing multi-layered feed-forward deep neural networks on integrated data that satisfies differential privacy, the classification performance of the integrated data can be enhanced [18]. Concurrently with the injection of noise into sensitive data, noise can also be introduced into classification models, such as Naive Bayes classifier. Consequently, employing individual privacy constraints can yield more robust privacy protection and improved data utility [19].

These methods facilitate the safeguarding of data within cloud services against potential privacy breaches. For data sharing on local services, various privacy-preserving algorithms based on differential privacy models have been developed, including techniques like dimensionality reduction using enhanced principal component analysis, or random projection, as well as segmentation of attributes values using spectral clustering [20–22]. These methods involve introducing noise to the modified dataset to achieve differential privacy protection. However, the changed attributes add to the complexity in data interpretation.

Instead of changing attributes and adding noise to achieve privacy protection, there were also studies that use the statistical transformation method such as weight of evidence and information value to implement initial data perturbation to the original dataset. After the completion of the first stage of data perturbation, a variety of privacy protection methods were applied to further improve the ability of privacy protection, such as privacy-chain based homomorphic encryption scheme, differential privacy scheme using Laplace mechanism, intuitionistic fuzzy Gaussian membership function and three-dimensional shearing [23–26]. These studies used multiple classifier models to evaluate the performance of perturbed datasets, demonstrating that the mining knowledge before the data transformation and after the data transformation was nearly equal.

Differential privacy mainly perturbs real data by adding noise, which can resist background knowledge attacks and differential attacks implemented by attackers, and is suitable for privacy protection in big data environment. Unlike data transformation, synthetic models can also

be used to generate new datasets instead of the release of real data. A hybrid approach that integrates differential privacy with synthetic data models has been developed. For instance, a probabilistic graphical model to understand the dependence between attributes in raw dataset has been combined with differential privacy to add specific noise to generate synthetic data that adheres to differential privacy standards. The resulting synthetic dataset retains the same attributes as the raw dataset, which facilitates data analysis and result interpretation. Recent discussions have focused on synthetic methods, such as differential privacy-based Bayesian network, and differential privacy-based Gibbs sampling [27–29].

For a set of attributes, a probabilistic graphical model can present the relationship between attributes, and Bayesian algorithms can reduce the computational difficulty and the dimensionality of the parameters. Zhang et al. [27] used a Bayesian network based on differential privacy to obtain a synthetic dataset, and they performed the performance comparison of the proposed model on four datasets. Specifically, they used greedy Bayesian algorithm to construct a Bayesian network, and they explored the choice of parameters for the Bayesian network and the form of noise. After that, they used ancestral sampling to sample the Bayesian network, which randomly selected a value of the ancestor and through noisy conditional probability decided the descendant. Yubin et al. [29] used Gibbs sampling based on differential privacy to obtain a synthetic dataset. Unlike in differential privacy-based Bayesian network where the marginal distributions incorporate random noise from Laplace distribution, they used Dirichlet prior perturbation to smooth out raw counts to satisfy differential privacy. Then, based on a seed, they used noisy conditional probability to sample every variable in turn.

A Bayesian network describes the relationship between a variable and its parent node through conditional probability distributions [30]. Markov Chain Monte Carlo (MCMC) sampling methods are often applied to theoretically guarantee that the sampled dataset can fit the full joint probability distribution, and Gibbs sampling converts complex multi-parameter problems into simple problems dealing with a single parameter at a time and theoretically improve the power of sampling and control computational burden [31, 32]. While two Bayesian synthetic algorithms based on differential privacy can produce synthetic datasets that uphold privacy protection, they may result in information loss, potentially leading to suspicious information in the subsequent data analysis process.

Balancing privacy protection with the utility of synthetic data is a crucial consideration in real-world data mining. From the literature review, the research gap can be found. The model based on data anonymization and data anatomization can protect sensitive data to a certain extent, but the generalization and suppression techniques affected the utility of data [8–10]. Considering that the complete dataset was more conducive to data analysis and result interpretation in data mining, the application of privacy protection model based on dimensionality reduction was limited [20–22]. Different from the multiple perturbations of the original dataset, it is also a worthy research direction to produce new synthetic dataset with noise by using the correlation between attributes [23–26]. An overview of the literature review is provided in Additional file 2.

Based on the literature review and research gap found in the studies, this study takes into account that the performance of Bayesian synthetic algorithms based on differential privacy is expected to vary based on data characteristics and the level of privacy protection required. Hence, comparing these synthetic algorithms across various scenarios is a valuable area of research that can offer insights for selecting appropriate methods in real-world data applications.

In this study, we used physical examination data obtained from a hospital health management center as a case study. We applied two Bayesian synthetic algorithms that combined different noise forms to generate synthetic datasets. Specifically, for algorithm 1, we used Bayesian network based on differential privacy (PrivBayes), with random noise sourced from a Laplace distribution. For algorithm 2, we added virtual samples as noise to each variable category to get noisy conditional probabilities, which were then sampled by Gibbs Sampler (PeGS). Additionally, we used random noise from a Laplace distribution to get noisy conditional probabilities and sampled by Gibbs Sampler for algorithm 3, referred to as DP-Gibbs in this study.

The main contributions of this study are listed following. First, we explored the combination model of different noise forms and Bayesian synthetic algorithms, and compared the existing algorithms to prove the effectiveness of the Bayesian model based on differential privacy. Second, we explored the balance between privacy protection and data utility under different levels of privacy protection, and realized the mining and application of a wider range of medical data under the requirements of privacy protection.

The rest of the study is organized as follows: Sect. 2 explains the materials and methods used in our study. Section 3 presents the results, encompassing the statistical similarity between synthetic datasets and the original dataset, privacy preserving capacity, machine learning performance, and a comparison of execution time. Section 4 discusses our findings and future directions. Section 5 provides a summary of this study.

Methods

Data pre-processing

The dataset used in this study comprised the physical examination data from Zhongda Hospital Affiliated to Southeast University, located in Nanjing, China, for the year 2021. The raw dataset initially contained variables with excessive missing values and duplicate personal records, which were removed. As the study was to evaluate three algorithms, the easy-to-understand and important indexes were filtered to be included as variables in the preprocessed dataset, forming the original dataset.

The original dataset consisted of 13,250 individual records and 9 variables, including *Age*, *Height*, *Weight*, *Sex*, *Time*, *Career*, *BMI*, *BP*, and *HPL*, with three continuous variables *Age*, *Height*, and *Weight*. In this study, the elbow method was utilized to determine the number of clusters for the continuous variables, and an additional file shows this in more detail [see Additional file 3].

After determining the number of clusters, the discretization of the continuous variables was then achieved by using the Bisecting K-Means Algorithm. Nine variables incorporated in the original dataset are shown in Table 1.

Table 1 Nine variables in physical examination data in 2021 from Zhongda hospital affiliated to Southeast university

Variable Name	Description	Category Values
Age	Age of the individual (years, mean (SD))	54.3 (4.2): 0; 28.4 (3.9): 1; 72.8 (6.9): 2; 40.9 (3.7): 3
Height	Height of the individual (m, mean (SD))	1.81 (0.04): 0; 1.56 (0.04): 1; 1.65 (0.03): 2; 1.73 (0.02): 3
Weight	Weight of the individual (kg, mean (SD))	93.4 (9.0): 0; 63.5 (3.4): 1; 76.0 (4.1): 2; 52.3 (3.8): 3
Sex	Sex of the individual	Male: 0; Female: 1
Time	Time of physical examination	Jan-Mar: 0; Apr-Jun: 1; Jul-Sep: 2; Oct-Dec: 3
Career	Occupation of the individual	Classified according to Industrial classification for national economic activities (GB/T 4754 – 2017), 18 levels: 0–17
BMI	Body mass index of the individual	Low weight: 0; Normal weight: 1; Overweight: 2; Obesity: 3
BP	The blood pressure level of the individual	Hypotension: 0; Normal BP: 1; Ideal BP: 2; Elevated BP: 3; Stage 1 hypertension: 4; Stage 2 hypertension: 5; Stage 3 hypertension: 6
HPL	The lipid level of the individual	Normal lipid level: 0; Hyperlipemia: 1

Differential privacy

The implementation of differential privacy allows for the sharing of datasets containing sensitive information while maintaining a specified level of privacy protection. It is assumed that there exists a dataset D_1, and a query operation F (e.g., counting, summing) performed on this dataset yielding result $F(D_1)$. When a record about individual x is added to dataset D_1, resulting in dataset D_2. The same query operation can be performed on D_2 to obtain $F(D_2)$. If $F(D_1)$ and $F(D_2)$ are indistinguishable, then no information about individual x is leaked.

Hence, differential privacy is formulated as follows:

$$\epsilon \geq ln\frac{Pr[M(D_1) \subseteq O]}{Pr[M(D_2) \subseteq O]} \quad (1)$$

Where randomized function M yields the query results after noise addition, O is the set of all possible query results for the dataset, and ε is privacy budget, which means the risk of privacy leakage. Function M is called ε-differentially private if Eq. (1) is satisfied. There are different methods to fulfill differential privacy, and the most frequently used differential privacy mechanisms for continuous data and discrete data are Laplace mechanism and exponential mechanism respectively [33–35]. Accordingly, the higher ε, the higher the risk of privacy leakage and the lower the level of protection of private information.

The overall workflow of the proposed framework for enhancing privacy protection of physical examination data is summarized as follows:

I. Data collection and preprocessing Raw physical examination data are collected and preprocessed by removing variables with excessive missing values and duplicates. Continuous variables are discretized using the Bisecting K-Means Algorithm.

II. Selection of synthetic data generation method The user selects one of the differential privacy-based Bayesian synthetic algorithms (PrivBayes, PeGS, or DP-Gibbs).

III. Model construction and noise injection The chosen algorithm analyzes the statistical relationships among variables and constructs either a Bayesian network or conditional probability distributions. Differential privacy is achieved by injecting noise via mechanisms such as the Laplace mechanism or virtual samples, depending on the algorithm.

IV. Synthetic data generation The privacy-preserving model is used to generate synthetic datasets that resemble the statistical properties of the original data, but with enhanced privacy protection.

V. Evaluation The synthetic data are evaluated for utility (statistical similarity, machine learning performance) and privacy (privacy preserving capacity).

Different differential privacy-based Bayesian synthetic algorithms are described in following sections, and a schematic illustration of this workflow is shown in Fig. 1.

Differential privacy-based bayesian network (PrivBayes)

Let the variables in the Bayesian network be X_1, X_2,..., X_n, and a complex full joint distribution can be decomposed into a product of conditional probability distributions for each variable:

$$P(X_1, ..., X_i) = \prod_{i=1}^{n} P\left(X_i | \pi\ (X_i)\right) \tag{2}$$

Where X_i is one of the set of variables, and $\pi(X_i)$ is the set of parent_node variables of X_i. By setting the number of variables in $\pi(X_i)$, different Bayesian networks can be generated using the greedy Bayesian algorithm.

This study developed a Bayesian network based on differential privacy (PrivBayes), which involved a three steps process. The algorithm is shown below.

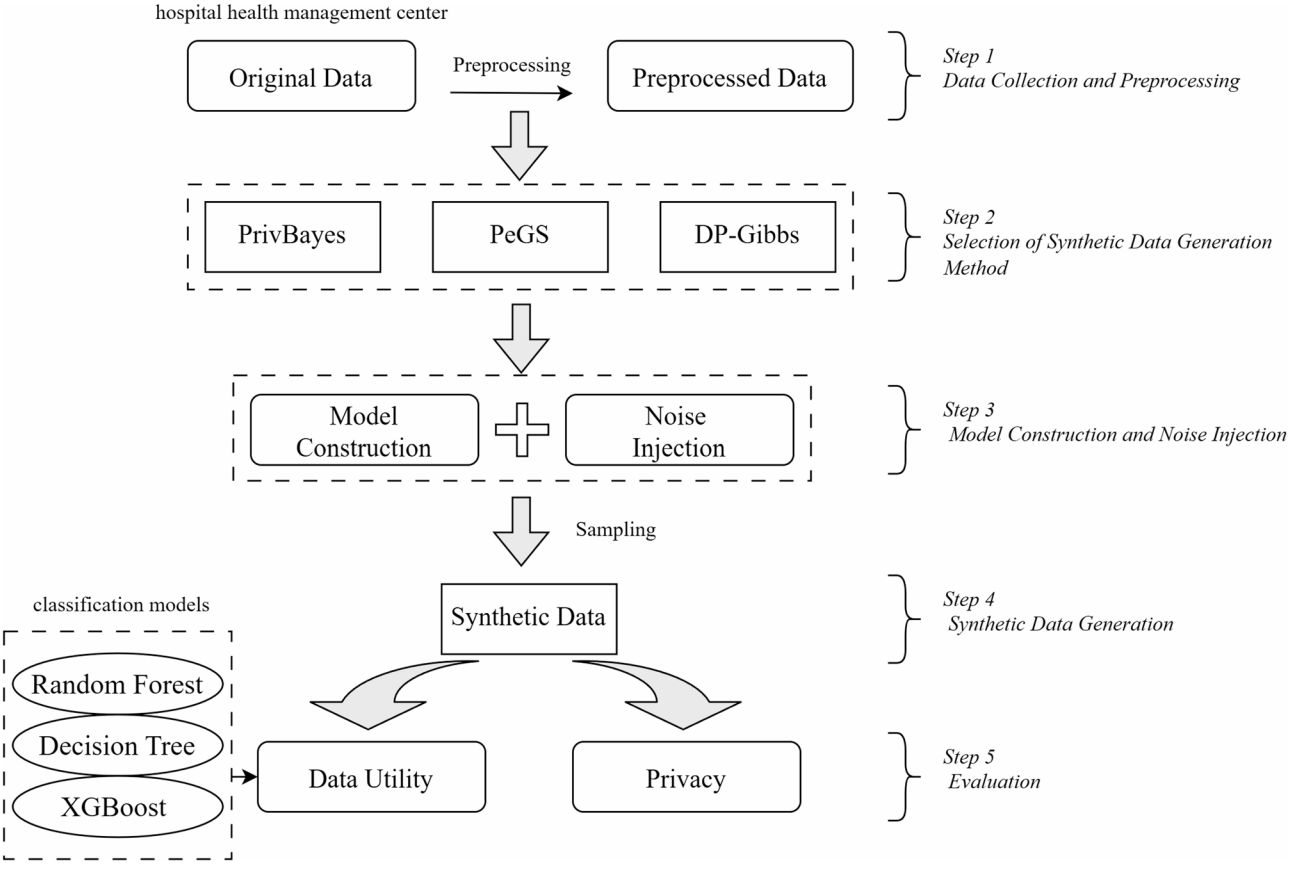

Fig. 1 Outline of the proposed work

Algorithm 1 (PrivBayes)
Generate synthetic dataset for physical examination dataset:

Input: Physical examination dataset D of size $n * d$, m-the number of variables in $\pi(X_i)$, privacy budget ε
Output: Synthetic dataset D_1 of size $n * d$
Step 1
1 Initialize $N = \Phi$, $V = \Phi$
2 Randomly select a variable X_1 from the set of variables. Add (X_1, V) to N and X_1 to V
3 For i =2 to d, i++:
  Initialize $\Omega = \Phi$
  For every variable $X_i \in X \setminus V$ and $\pi(X_i) \in V$ (the maximum number of variables in $\pi(X_i)$ is m). Add $(X_i, \pi(X_i))$ to Ω
  Set privacy budget ε_1 and use exponential mechanism to select a $(X_i, \pi(X_i))$ from Ω
  Add $(X_i, \pi(X_i))$ to N and X_i to V
4 Get N (m-degree Bayesian network)
Step 2
5 Initialize $P = \Phi$
6 For $i = m+1$ to d, i++:
  Construct the joint distribution $\Pr[X_i, \pi(X_i)]$ of X_i
  Set privacy budget ε_2 and use Laplace mechanism to add noise to $\Pr[X_i, \pi(X_i)]$
  Extract $\Pr^*[X_i \mid \pi(X_i)]$ from $\Pr^*[X_i, \pi(X_i)]$ and add $\Pr^*[X_i \mid \pi(X_i)]$ to P
7 For i =1 to m, i++:
  Extract $\Pr^*[X_i \mid \pi(X_i)]$ from $\Pr^*[X_{m+1}, \pi(X_{m+1})]$ and add $\Pr^*[X_i \mid \pi(X_i)]$ to P
8 Get P (noisy conditional probability distributions)
Step 3
9 Initialize $D_1 = \Phi$
10 For i =1 to n, i++:
  For j =1:
    Use $\Pr[X_j]$ to select a value of X_j
  For j =2 to d, j++:
    Use P to select a value of X_j
  Add i-th record to D_1
11 Return D_1 (synthetic dataset)

The degree of the Bayesian network was first determined. The ratio of the average scale of the amount of information to the average scale of noise should be large. The average scale of the amount of information can be denoted as $1/2^{m+1}$, and the average scale of noise can be denoted as $2^*(d\text{-}m)/(n^*\varepsilon_2)$. Hence, to maintain a large ratio, as privacy budget increases, the value of m should become larger. In this study, we set m= {1, 2, 3, 4, 5}, when ε= {0.5, 1, 3, 5, 10}. An example of a Bayesian network is shown in Fig. 2.

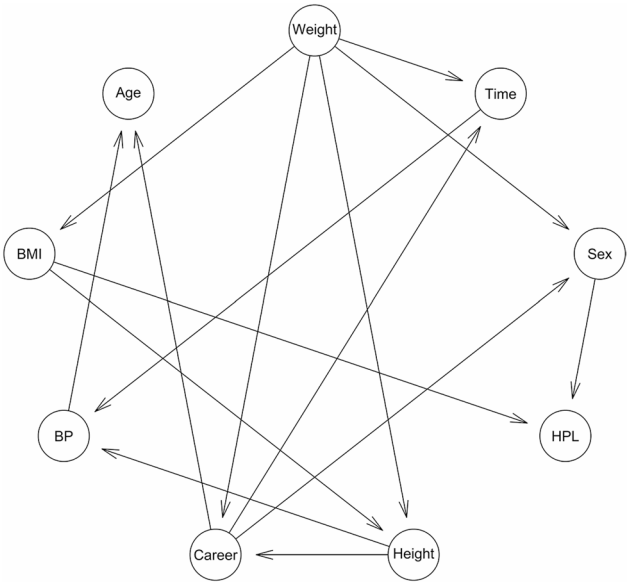

Fig. 2 An example of 2-degree Bayesian network

To satisfy differential privacy, we used exponential mechanism to build a score function based on the mutual information between X_i and $\pi(X_i)$. A probability value was obtained for each $(X_i, \pi(X_i))$, by which we selected a variable to be included in the Bayesian network. Meanwhile, we used Laplace mechanism, and the random noise that satisfied the Laplace distribution was added to the joint distribution of each variable in the Bayesian network.

A Bayesian network satisfying the total privacy budget parameter ε, the sum of ε_1 and ε_2, was utilized for sampling to obtain a synthetic dataset of the same size as the original dataset. Let β be the allocation ratio of ε, where $\epsilon_1 = \beta \times \epsilon, \epsilon_2 = (1-\beta) \times \epsilon$. With reference to [13], the value of β should be within the range of [0.2,0.5], when β was 0.3, the utility of synthetic datasets was better. In this study, $\beta = 0.3$ was set.

Differential privacy-based Gibbs sampling (PeGS)

Considering differential privacy and Gibbs sampling, for a random seed s, a synthetic record x can be synthesized by a probability, which is denoted as follows:

$$Pr_{D_1,\alpha}(x|s) = \prod_{i=1}^{d} Pr_{D_1,\alpha}(x_i|x_{1:(i-1)}, s_{(i+1):d}) \quad (3)$$

Where $x_{1:0}$ and $s_{(d+1):d}$ are null values. For two datasets D_1 and D_2, differing by one record, (1) can be rewritten as follows:

$$\epsilon \geq ln\frac{Pr_{D_1,\alpha}(x|s)}{Pr_{D_2,\alpha}(x|s)} = ln\frac{\prod_{i=1}^{d} Pr_{D_1,\alpha}(x_i|x_{1:(i-1)}, s_{(i+1):d})}{\prod_{i=1}^{d} Pr_{D_2,\alpha}(x_i|x_{1:(i-1)}, s_{(i+1):d})} \quad (4)$$

We used Dirichlet prior perturbation to smooth out raw counts to satisfy differential privacy, which meant α virtual samples were added to each category of each variable. For i-th variable, we can get:

$$Pr_{D_1,\alpha}(X_i = j|X_{-i}) = \frac{n_{ij} + \alpha}{N_{X_{-i}} + C_i\alpha} \quad (5)$$

Where N_{X-i} is the count of records that have the same conditional variable values. The C_i is the category of values for variable X_i, and n_{ij} is the count of the j-th category within N_{X-i}. In this study, the relationship between α and ε is $\alpha = 1 / (\exp(\varepsilon/d) - 1)$.

The algorithm of PeGS is outlined below.

Algorithm 2 (PeGS)
Generate synthetic dataset for physical examination dataset:

Input: Physical examination dataset D of size $n * d$, m-the number of conditional variables for X_i, privacy budget ε
Output: Synthetic dataset D_2 of size $n * d$
Step 1
1 Initialize $\Omega = \Phi$
2 For i =1 to d, i++:
Initialize $C(X_i) = \Phi$
For $X_i \in X$ and every variable $V_j \in X \setminus X_i$ ($j \in$ [1, d-1]), calculate the mutual information of X_i and V_j
Select m variables with the maximum mutual information from V, and add to $C(X_i)$
Use feature hashing [36] to disintegrate the joint distribution of (X_i, $C(X_i)$) to form statistical blocks, and add to Ω
3 Get Ω (some statistical blocks and each statistical block has the same values with (X, $C(X)$))
Step 2
4 Initialize $P = \Phi$
5 For i =1 to d, i++:
Sum the count of records in some statistical blocks, and construct the joint distribution $\Pr[X_i, C(X_i)]$ of X_i
Set privacy budget ε and $\alpha = 1 / (\exp(\varepsilon/d) - 1)$
Add α to $\Pr[X_i \mid C(X_i)]$ and add $\Pr^*[X_i \mid C(X_i)]$ to P
6 Get P (noisy conditional probability distributions)
Step 3
7 Initialize $D_2 = \Phi$
8 For i =1 to n, i++:
For i-th seed $s \in D$
For j =1 to d, j++:
$x_j \leftarrow P(X_j \mid x_{1:(j-1)}, s_{(j+1):d})$
null values $x_{1:0}$ and $s_{(d+1):d}$
Add i-th record x to D_2
9 Return D_2 (synthetic dataset)

Differential privacy-based Gibbs sampling (DP-Gibbs)
For DP-Gibbs, we used Laplace distribution to generate random noise to perturb real conditional probability distributions, and then used Gibbs sampling to generate synthetic dataset. As algorithm 3, DP-Gibbs modifies the step 2 in Algorithm 2 appropriately.

Algorithm 3 (DP-Gibbs)
Generate synthetic dataset for physical examination dataset:

Input: Physical examination dataset D of size $n * d$, m-the number of conditional variables for X_i, privacy budget ε
Output: Synthetic dataset D_3 of size $n * d$
Step 1 - Get Ω
Step 2
4 Initialize $P = \Phi$
5 For i =1 to d, i++:
Sum the count of records in some statistical blocks, and construct the joint distribution $\Pr[X_i, C(X_i)]$ of X_i
Set privacy budget ε and use Laplace mechanism to add noise to $\Pr[X_i, C(X_i)]$
Extract $\Pr^*[X_i \mid C(X_i)]$ from $\Pr^*[X_i, C(X_i)]$ and add $\Pr^*[X_i \mid C(X_i)]$ to P
6 Get P (noisy conditional probability distributions)
Step 3 - Return D_3

In three steps, the synthetic datasets result by three algorithms (PrivBayes, PeGS, or DP-Gibbs) have been done, and Table 2 presents the values of some variables when ε was 0.5.

Assume that the quasi-identifier of this record is known to the attacker (e.g., male in career category 4). When the record is incorporated into the database, it forms a new raw dataset, and is published by synthetic algorithms. This record will be converted into a new instance (such as a male in career category 17), and the attacker will not be able to obtain the corresponding sensitive information from the existing background knowledge.

Simulation study setting

The implementation of PeGS and DP-Gibbs relied on the construction of conditional probability distributions. The number of conditional variables for X_i was set to 3, and the variables that realize the joint distribution after dimensionality reduction were 4 out of 9.

Five levels of privacy budget were established, with ε= {0.5, 1, 3, 5, 10}. The original dataset was employed to generate a synthetic dataset with the same number of records as the original dataset using PrivBayes, PeGS, or DP-Gibbs. To address uncertainty, each algorithm was subjected to repeated experiments, with 100 repetitions determined based on a pilot study. The average results were then reported.

All data processing and data analysis were conducted using R 4.2.2 software.

Table 2 Synthetic data generated by differential privacy-based bayesian algorithms

Operation	Age	Sex	Career	BP	HPL
Real Values	1	0	4	1	0
PrivBayes	0	0	6	2	0
PeGS	1	0	10	1	1
DP-Gibbs	3	0	17	2	0

Results

Evaluation metrics

This section highlights the results of proposed algorithms by conducting tests using physical examination data. The performance of the three synthetic algorithms was evaluated through comparisons of statistical similarity, privacy preserving capacity, machine learning performance, and execution time. And the result was also compared with the existing algorithm Private-PGM [37].

Kullback Leibler (KL) divergence [38] can be utilized to calculate the similarity between the distribution of two variables. From the perspective of information loss, if two distributions are more similar, then KL divergence will be smaller. And KL divergence is always non-negative. In this study, the marginal probability mass function of each variable in the synthetic dataset and the original dataset was calculated independently, and then the KL divergence was calculated. The statistical similarity between two datasets was measured by summing the KL divergence of each pair of variables and normalizing it by 1/(1 + KL divergence).

Privacy Preserving Capacity (PPC) plays an important role in data security. In this study, the ability of several algorithms to protect sensitive data under different privacy levels was measured by calculating the variance measure of the original dataset and the synthetic dataset.

And higher PPC means better privacy benefits for the synthetic algorithm.

In terms of machine learning performance, we used Random Forest, Decision Tree, and XGBoost for analysis. These classifiers [39] used the remaining 8 variables to predict target variables specified, hypertension. Each classifier used 70% of the dataset for training and 30% for testing. We used precision, recall, F1-score, misclassification rate, kappa-coefficient and AUC-score (Area Under the ROC Curve) to illustrate the machine learning performance of synthetic datasets as well as the original dataset.

Statistical similarity

The KL divergence score of different synthetic datasets generated and the original dataset are shown in Table 3.

The KL divergence score closer to 1 indicates a higher similarity between two datasets. Given the privacy budget, it can be found that synthetic datasets generated by PeGS and DP-Gibbs exhibited greater similarity to the original dataset compared to those generated by PrivBayes. Compared with Private-PGM, the KL divergence score of PeGS and DP-Gibbs show a similar upward trend as the level of privacy protection decreases. When ε was set to 0.5, 1, 3, 5, 10, DP-Gibbs exhibited progressively increasing KL divergence score of 0.637, 0.799, 0.950, 0.973, 0.987. When the privacy budget increased, the statistical similarity of these algorithms also improved. However, PrivBayes has a lower KL divergence score, with fluctuations lacking a clear pattern. In the context of our physical examination data, PeGS and DP-Gibbs were able to preserve more information compared to PrivBayes.

Privacy preserving capacity

This study used Privacy Preserving Capacity to measure how well the proposed algorithms protect data privacy compared to the state-of-art algorithm. Higher PPC means that the algorithm achieves better privacy protection. Taking the blood pressure variable in physical examination data as an example, Table 4 shows the PPC of four algorithms.

From Table 4, it can be found that when the privacy budget is small, both PeGS and DP-Gibbs have higher PPC, 4.354 and 4.686 respectively, indicating that the Gibbs sampling-based synthetic algorithm is superior to the Bayesian network-based algorithm in privacy protection. Compared with the existing algorithm, which has a stable PPC when the privacy budget changes, PeGS and DP-Gibbs have better PPC within a given range of privacy levels, indicating that PeGS and DP-Gibbs are indeed better than Private-PGM in privacy protection.

Table 3 Results of measuring statistical similarity between original and synthetic datasets (KL divergence score)

Algorithms	Privacy Budget ε				
	0.5	**1**	**3**	**5**	**10**
PrivBayes	0.386	0.412	0.412	0.388	0.352
PeGS	0.669	0.807	0.950	0.977	0.990
DP-Gibbs	0.637	0.799	0.950	0.973	0.987
Private-PGM	0.691	0.798	1.000	1.000	1.000

Table 4 PPC comparison for four synthetic datasets

Algorithms	Privacy Budget ε				
	0.5	**1**	**3**	**5**	**10**
PrivBayes	2.063	2.277	2.582	3.596	4.512
PeGS	4.354	3.636	2.675	2.365	2.158
DP-Gibbs	4.686	3.859	2.644	2.416	2.215
Private-PGM	2.012	2.001	2.001	2.000	2.000

Machine learning performance

The quality of a synthetic dataset can be gauged by how closely its machine learning performance aligns with that of the real dataset.

Precision is a measure of how accurate the positive predictions of a model are, and Recall measures the ability of a model to capture all instances of the positive class. F-Measure provides a single score that balances both the concerns of precision and recall in one number. Table 5 presents the Precision, Recall, and F1-score of synthetic datasets (under various privacy budgets) and real dataset using different classifiers.

Misclassification rate is calculated as the number of total incorrect predictions divided by the total number of predictions. Kappa Coefficient is an indicator used for consistency test and can be used to measure the effect of classification. Table 6 presents the Misclassification Rate and Kappa Coefficient of synthetic datasets (under various privacy budgets) and real dataset.

Considering the Precision, Recall and F1-score comprehensively, the results in the Table 5 show that the machine learning performance of the real dataset is better when the model is trained by Random Forest (*Precision = 0.571*, *Recall = 0.346*, *F1-score = 0.431*). When applying Random Forest for classification prediction of synthetic datasets, under different privacy budgets, DP-Gibbs has more consistent Precision and F1-score with real dataset. When ε was 10, DP-Gibbs had a Precision of 0.540 and F1-score was 0.428, which was better than other algorithms. With the increase of privacy budgets, the Recall of DP-Gibbs gradually approaches the performance of real dataset.

From Table 6, it can be found that when applying multiple Classifier Models, the Misclassification Rate and Kappa Coefficient of DP-Gibbs gradually approach the level of real dataset with the increase of privacy budget, and the consistency of its Kappa Coefficient with the real value is better than other algorithms, showing good

Table 5 Precision, recall, F1-score of original and four synthetic datasets

Algorithms	Classifier Models	Precision						Recall						F1-score					
		$\varepsilon=0.5$	$\varepsilon=1$	$\varepsilon=3$	$\varepsilon=5$	$\varepsilon=10$	∞	$\varepsilon=0.5$	$\varepsilon=1$	$\varepsilon=3$	$\varepsilon=5$	$\varepsilon=10$	∞	$\varepsilon=0.5$	$\varepsilon=1$	$\varepsilon=3$	$\varepsilon=5$	$\varepsilon=10$	∞
PrivBayes	RF	0.458	0.577	0.630	0.684	0.697	0.571	0.153	0.279	0.411	0.489	0.519	0.346	0.230	0.371	0.498	0.570	0.595	0.431
	DT	0.624	0.629	0.615	0.595	0.598	0.609	0.093	0.175	0.249	0.195	0.172	0.249	0.162	0.274	0.355	0.294	0.267	0.353
	XGB	0.621	0.641	0.640	0.631	0.610	0.595	0.097	0.227	0.337	0.323	0.304	0.249	0.168	0.335	0.441	0.427	0.406	0.351
PeGS	RF	0.465	0.483	0.523	0.528	0.549		0.328	0.334	0.345	0.346	0.325		0.385	0.395	0.415	0.418	0.408	
	DT	0.521	0.547	0.587	0.584	0.605		0.130	0.277	0.260	0.270	0.267		0.208	0.368	0.360	0.369	0.371	
	XGB	0.517	0.538	0.581	0.580	0.595		0.167	0.260	0.299	0.319	0.316		0.252	0.350	0.395	0.411	0.413	
DP-Gibbs	RF	0.620	0.592	0.546	0.557	0.540		0.477	0.449	0.386	0.394	0.355		0.539	0.510	0.453	0.461	0.428	
	DT	0.635	0.632	0.600	0.589	0.603		0.368	0.338	0.320	0.306	0.273		0.466	0.441	0.417	0.402	0.376	
	XGB	0.655	0.634	0.602	0.585	0.589		0.434	0.422	0.377	0.360	0.341		0.523	0.506	0.464	0.445	0.432	
Private-PGM	RF	0.520	0.527	0.531	0.534	0.534		0.233	0.252	0.267	0.265	0.267		0.321	0.340	0.355	0.354	0.356	
	DT	0.616	0.616	0.618	0.615	0.618		0.207	0.243	0.262	0.265	0.263		0.310	0.349	0.368	0.370	0.369	
	XGB	0.612	0.614	0.617	0.614	0.616		0.210	0.244	0.263	0.265	0.263		0.313	0.349	0.369	0.370	0.369	

Table 6 Misclassification rate, kappa coefficient of original and four synthetic datasets

Algorithms	Classifier Models	Misclassification Rate (%)						Kappa Coefficient					
		$\varepsilon=0.5$	$\varepsilon=1$	$\varepsilon=3$	$\varepsilon=5$	$\varepsilon=10$	∞	$\varepsilon=0.5$	$\varepsilon=1$	$\varepsilon=3$	$\varepsilon=5$	$\varepsilon=10$	∞
PrivBayes	RF	27.8	26.2	23.7	22.2	21.1	24.6	0.109	0.229	0.352	0.427	0.456	0.286
	DT	26.0	25.8	25.9	28.2	28.1	24.5	0.099	0.174	0.228	0.170	0.154	0.234
	XGB	26.0	25.0	24.3	26.0	26.5	24.1	0.102	0.222	0.305	0.281	0.259	0.287
PeGS	RF	34.0	30.8	27.6	26.7	25.3		0.160	0.198	0.245	0.254	0.261	
	DT	32.0	28.6	26.3	25.6	24.4		0.090	0.208	0.225	0.235	0.247	
	XGB	32.0	29.0	26.1	25.3	24.1		0.112	0.192	0.250	0.267	0.278	
DP-Gibbs	RF	28.1	27.9	26.6	24.1	25.7		0.342	0.320	0.283	0.312	0.271	
	DT	29.1	27.9	25.4	23.8	24.6		0.284	0.277	0.273	0.272	0.249	
	XGB	27.4	26.7	24.8	23.5	24.3		0.341	0.334	0.313	0.307	0.291	
Private-PGM	RF	26.1	26.0	25.8	25.8	25.8		0.188	0.203	0.215	0.216	0.216	
	DT	24.8	24.3	24.0	24.1	24.1		0.203	0.233	0.249	0.249	0.249	
	XGB	26.6	26.3	26.0	26.1	26.1		0.204	0.233	0.249	0.249	0.249	

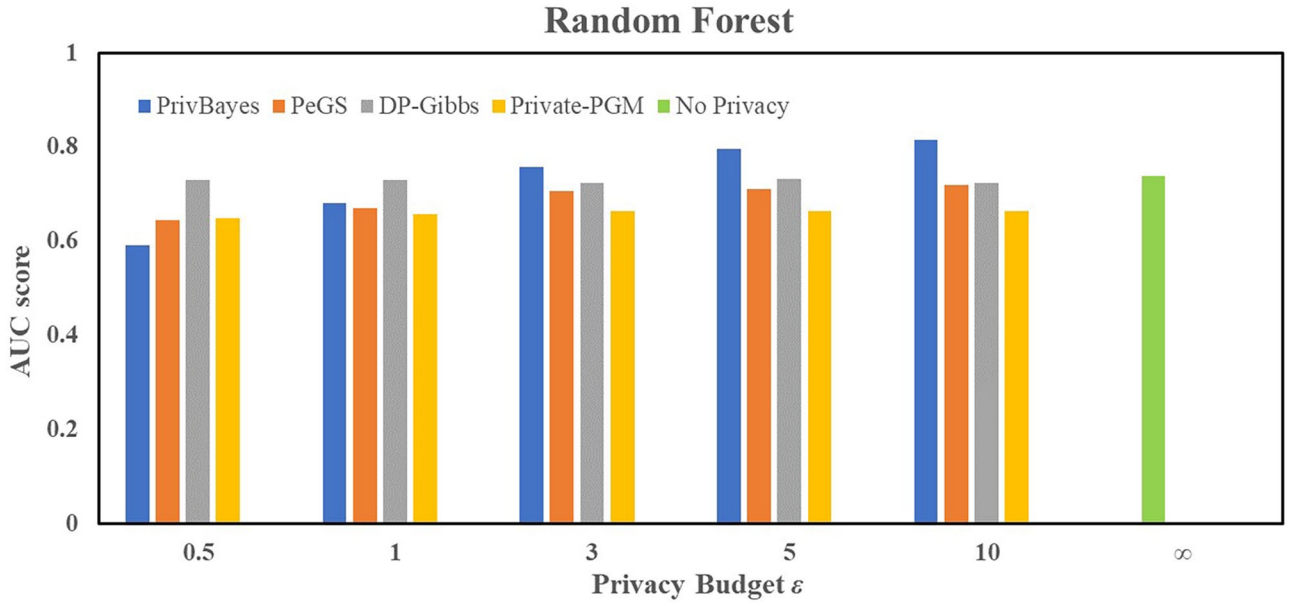

Fig. 3 Results of AUC-score by training Random Forest on original and synthetic datasets

data availability. Applying XGBoost to the real dataset obtained a Kappa Coefficient of 0.287. When ε was set to 0.5, 1, 3, 5, 10, DP-Gibbs exhibited progressively similar Kappa Coefficient of 0.341, 0.334, 0.313, 0.307, 0.291.

The AUC-score considers both False Positive Rate and True Positive Rate, which can measure the performance of Classifier Models. Visualize the results of AUC-score as shown in Fig. 3. and Fig. 4.

From two figures, it can be seen that DP-Gibbs shows unique advantages in both Random Forest and XGBoost, and its AUC-scores are close to the true values. The true value of AUC-score via Random Forest was 0.738, and DP-Gibbs presented the AUC-scores of 0.729, 0.729, 0.723, 0.731, 0.723, with ε was set to 0.5, 1, 3, 5, 10. The true value via XGBoost was 0.777, and DP-Gibbs presented the AUC-scores of 0.765, 0.762, 0.764, 0.776, 0.768.

Obviously, the change trend of machine learning performance of DP-Gibbs is related with the change of its Privacy Preserving Capacity. With the improvement of data utility of synthetic algorithm, the degree of privacy protection gradually decreases. Under the requirement

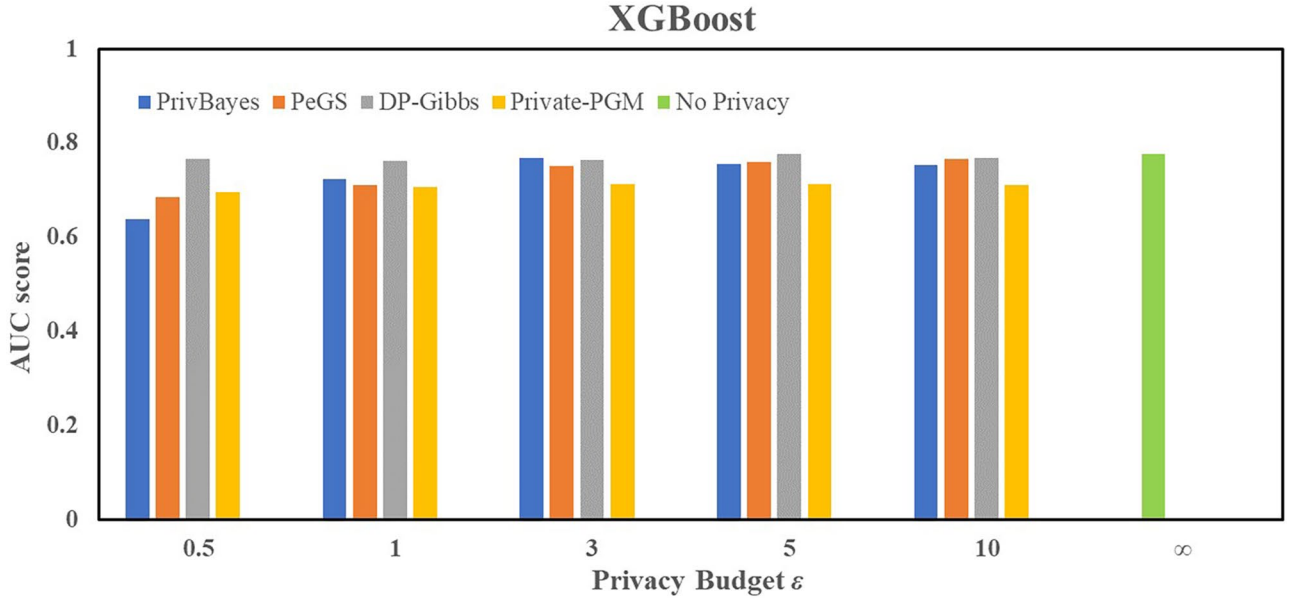

Fig. 4 Results of AUC-score by training XGBoost on original and synthetic datasets

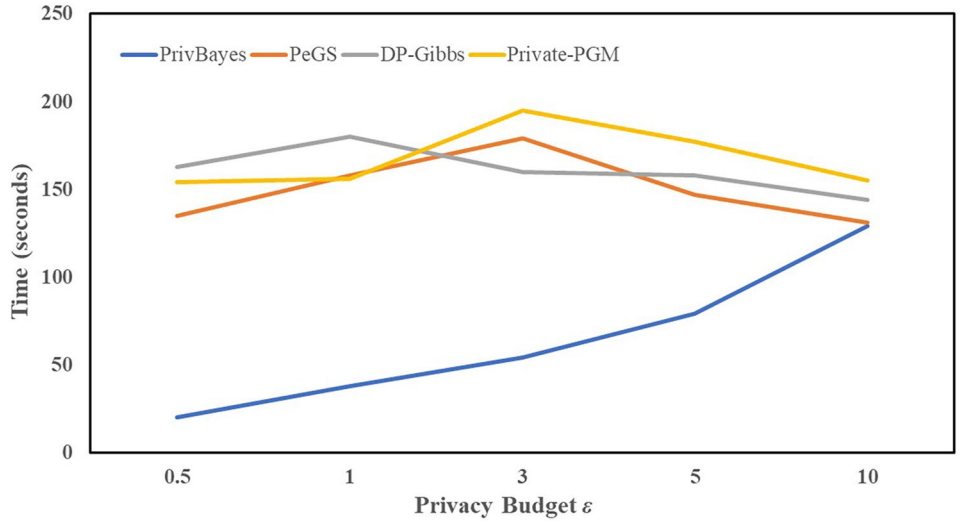

Fig. 5 The execution time of four algorithms for generating synthetic datasets

of privacy budgets, DP-Gibbs exhibits similar statistical outcomes to the original dataset, which satisfies a certain Privacy Preserving Capacity.

Execution time comparison

The simulations were run on laptop with 8 CPU cores and 16 GB RAM. The realization of three algorithms was done by R 4.2.2 software. The time taken by each algorithm to generate synthetic datasets under a given privacy budget was recorded individually. Figure 5. presents the comparison of execution time among these algorithms.

For time complexity, the core computational cost of PrivBayes arises from constructing the Bayesian network via the greedy algorithm and performing ancestral sampling. The network construction has complexity of O(d × m × n), where *d* is the number of variables, *m* the maximum parents per node, and *n* the number of records. Sampling is linear in *n* and *d*. Both PeGS and DP-Gibbs require the computation of conditional distributions using mutual information and block counting, followed by Gibbs sampling. The time complexity of conditional distribution construction is O(d × m × n). Gibbs sampling requires O(n × d × s) time, where *s* is the number of sampling iterations/seeds. The Private-PGM involves fitting graphical models to the data, with computational cost depending on the model order and dataset size, typically $O(n \times d^k)$, where *k* is the model's tree-width or maximum degree.

For space complexity, all methods require storing the dataset (O(n × d)) and the computed model parameters (e.g., conditional probabilities, sufficient statistics), which may additionally require O(d × M), where *M* is the maximum number of possible joint configurations for parents of a variable. DP-Gibbs and PeGS use block-based data structures for fast sampling, resulting in higher memory consumption compared to PrivBayes, in practice.

As shown in Fig. 5, DP-Gibbs and PeGS have higher execution times compared to PrivBayes. Private-PGM also demonstrates increased computation time for large datasets. Regarding space usage, DP-Gibbs and PeGS require more memory due to the storage of noisy conditional probabilities for all blocks, especially as the number of variables and their cardinality increase.

Discussion

Taking physical examination data as an example, this study evaluated synthetic datasets generated by three Bayesian synthetic algorithms based on differential privacy, and compared them with existing algorithms under specified privacy risk to further demonstrate the effectiveness of the proposed algorithm. Through statistical similarity, Privacy Preserving Capacity, and machine learning performance comparisons of synthetic datasets, this study found that the PPC of DP-Gibbs decreased with the increase of privacy budget, while the machine learning performance was closer to the performance of the real dataset. Under a given privacy budget, DP-Gibbs can achieve higher level of privacy protection and excellent machine learning performance compared with existing algorithms.

The DP-Gibbs may provide useful address on the trade-off between privacy protection and data utility. The superior performance of the DP-Gibbs framework in balancing privacy protection and data utility arises from several key methodological features: initially, differential privacy quantifies the level of privacy leakage, facilitating data sharing under specified privacy protection requirements. Secondly, in DP-Gibbs, we selected conditional

variables based on mutual information between variables, which improves data utility. In contrast to PrivBayes and Private-PGM, DP-Gibbs uses Gibbs sampler for iteratively sampling each variable, circumventing challenges like model selection and parameter estimation inherent in parametric modeling. Unlike the approach of adding noise through virtual samples, as employed by the PeGS, the Laplace mechanism enriches the data through the introduction of noise. The DP-Gibbs framework's methodological innovations, especially the integration of Gibbs sampling with direct and tunable noise injection, enable it to surpass prior state-of-the-art approaches in achieving both strong privacy protection and high utility in synthetic data generation.

The main contributions of this study are listed following. First, we explored the combination model of different noise forms and Bayesian synthetic algorithms, and compared the existing algorithms to prove the effectiveness of the Bayesian model based on differential privacy. Compared with changing attributes or data anonymization to achieve privacy protection, the idea based on synthetic algorithm can provide a more suitable dataset and realize the easy interpretation of statistical results. Differential privacy is based on strict mathematical definition to quantify the level of privacy protection, which is effectively applicable to big data application scenarios with strict privacy protection requirements. Second, we explored the balance between privacy protection and data utility under different levels of privacy protection. Privacy budgets take into account the sensitivity and noise of the data. In this study, five levels of privacy budget were set to evaluate the trade-offs of privacy protection and data utility under different Settings through Privacy Preserving Capacity and machine learning performance. As a proposed algorithm, DP-Gibbs generates synthetic datasets under different privacy standards, and the correlation between its privacy protection and data utility can be found. That is, a high privacy budget will cause a decrease in the level of Privacy Preserving Capacity, but improve the availability of data. Therefore, compared with the existing algorithm; by obtaining prior information and setting an appropriate privacy budget, DP-Gibbs can achieve the optimal collection of data information under a higher degree of privacy protection. Third, this study realized the mining and application of a wider range of medical data under the requirements of privacy protection. We focused on physical examination data. Compared with clinical data or outpatient data, physical examination data contains more useful information that has not yet been discovered and may be helpful for health management and disease prevention. This study provides an implementable privacy-preserving algorithm DP-Gibbs, which can generate synthetic datasets under specified privacy protection requirements for data mining of physical examination data rich in health information. This algorithm might assist healthcare professionals in accurate diagnoses and tailored guidance for the daily health management of specific populations.

It is worth further stating that, with the increase of physical examination data, health care managers have a more novel way to understand the health status of the vast majority of people, and possible undetected disease patterns. Our study concentrates on the health information of seemingly healthy individuals to achieve a comprehensive and professional assessment of the overall population's health status. However, physical examination data has the characteristics of large amount of data, various indicators, relatively few available indicators, and complete data also contains a lot of sensitive information. When the physical examination data is improperly used, it will cause the problem of meaningless statistical results and privacy disclosure. How to deal with physical examination data effectively to avoid privacy disclosure and contribute to health analysis is a research direction worthy of researchers' attention.

In order to make full use of physical examination data, one way is to achieve privacy protection through the release of noisy synthetic datasets by hospital health management center. There are also some challenges in generating synthetic datasets. Constructing an exact joint probability distribution can lead to divulging detailed original data and increase computation load. Conversely, converting the joint distribution into a product of univariate marginal distributions reduces the utility of synthetic data and overlooks crucial information from the original data's joint distribution. The determination of the number of conditional variables for each variable requires balancing computational complexity and appropriate conditional assumptions. Therefore, exploring adjustments to the number of conditional variables (denoted as "m") is essential for evaluating the future availability of the synthetic dataset generated by DP-Gibbs.

In addition to the construction of the model, the quality of the original dataset itself also affects the generation of the synthetic dataset. When some important indicators of physical examination data have very few positive values, imbalanced datasets will reduce the analysis of this indicator when constructing synthetic datasets, resulting in the loss of important information contained in physical examination data. Utilizing resampling-based data-augment techniques, such as Oversampling and Undersampling, may be a suitable solution. In addition to the impact of imbalanced datasets on the generation of synthetic datasets, when unrelated indicators are used as condition variables for data sampling, the information between indicators will be exaggerated, resulting in poor model performance. Therefore, it is worth paying

attention to the proper processing of physical examination data and its synthesis and publication.

There are some limitations in this study. The proposed DP-Gibbs synthetic data generation framework is currently optimized for tabular categorical data with moderate dimensionality. As the number of variables and parent set size increases, both computational and memory costs (for conditional probability table estimation and Gibbs sampling) rise substantially. While discretization enables the inclusion of continuous data, direct modeling of high-cardinality or mixed-type data may require either dimensionality reduction, hierarchical modeling, or hybrid Bayesian structures. Ultimately, model scalability and accuracy depend on careful design choices regarding network complexity, privacy budget allocation, and the adoption of emerging scalable methods. Future research will address extensions to support large-scale and more structured healthcare datasets (e.g., by integrating variational Bayesian approaches, parallelism, or advanced DP techniques) to further enhance modeling fidelity while preserving rigorous privacy guarantees.

Additionally, in this study, all continuous variables are discretized prior to modeling to enable the use of categorical Bayesian networks and to simplify the application of differential privacy mechanisms. While discretization is a common strategy, it unavoidably leads to information loss - particularly when the original variables exhibit subtle patterns, outliers, or non-uniform distributions relevant for downstream analysis. Recent advances in the field offer alternative methods for handling continuous variables under differential privacy with reduced information loss. For example, approaches based on DP-variational autoencoders (DP-VAE) or DP-generative adversarial networks (DP-GAN) can natively support both discrete and continuous features and often produce higher utility synthetic data [40, 41]. Integrating these state-of-the-art approaches into our framework represents a promising direction to reduce information loss while retaining differential privacy guarantees. In particular, developing differentially private hybrid Bayesian networks or leveraging DP deep generative models could allow direct, high-utility modeling of mixed-type healthcare data. We plan to investigate these extensions in future research.

Dimensionality reduction (such as principal component analysis, random projection, and feature selection [20–22]), when carefully applied under differential privacy, can enhance privacy while maintaining critical data structures—but also risks removing important variables, impacting utility. Future work will explore the integration of advanced and differentially private dimensionality reduction methods (e.g., DP-PCA, DP-autoencoders, DP-feature selection [42, 43]), to optimize privacy-utility trade-offs in synthetic healthcare data applications.

Finally, addition of formal statistical tests to quantify significance among algorithmic trade-offs is beyond the current scope but represents an important avenue for future work. Our present comparative analysis is grounded in repeated empirical trials and visualized trends, providing practical insights for real-world deployment.

Conclusion

As a result of increased health awareness, more and more physical examination data are generated, which contain useful information that has not yet been unearthed. There are a lot of indicators and sensitive information in physical examination data, which makes it difficult to make full use of it, resulting in the loss of health knowledge. In this study, a Bayesian synthetic algorithm based on differential privacy, DP-Gibbs, is proposed. The data manager first models the original dataset, constructs the available probability distributions, injects noise into the distributions through differential privacy, and then generates the synthetic dataset with the noised conditional probability distributions. Compared with the existing algorithms, DP-Gibbs proposed in this study has certain advantages and can achieve privacy protection, that is, it can meet a certain privacy preserving capacity and maintain high data utility at the same time, which is conducive to the interpretation of statistical results of datasets. Hence, DP-Gibbs can aid clinical researchers in conducting more in-depth data mining on physical examination data, thereby assessing medical research insights on diseases and the health status of a broader populations.

In future work, DP-Gibbs can be applied to more complicated physical examination data including both continuous variables and categorical variables, to capture more information through high-dimensional network structure, and at the same time to explore solution to high computing burden caused by the complicated network, like variational inference technique.

Abbreviations

PrivBayes	Differential privacy-based Bayesian network
PeGS	Differential privacy-based Gibbs sampling (virtual samples)
DP-Gibbs	Differential privacy-based Gibbs sampling (Laplace distribution)
PPC	Privacy Preserving Capacity
F1-score	The harmonic mean of precision and recall
AUC	Area under the curve
ROC	Receiver operating characteristic curve

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12911-025-03109-1.

Supplementary Material 1: Peer review

Supplementary Material 2

Supplementary Material 3

Acknowledgements
Not applicable.

Author contributions
WL.Z. and XJ.Y. provided conceptual and statistical guidance for all aspects of the work. WL.Z. performed the statistical analysis, prepared figures and tables, and drafted the initial manuscript. XJ.Y. and DP.J. provided guidance for statistical method and edited the manuscript extensively. R.L. and XY.Z. were in charge of data collection, provided guidance for this work. All the authors approved the final version of the manuscript.

Peer review
The peer review reports can be found at https://doi.org/10.1186/s12911-025-3109-1.

Funding
Not applicable.

Data availability
The data that support the findings of this study are available from Zhongda Hospital Affiliated to Southeast University but restrictions apply to the availability of these data, which were used under license for the current study, and so are not publicly available. Data are however available from the authors upon reasonable request and with permission of Zhongda Hospital Affiliated to Southeast University.

Declarations

Ethics approval and consent to participate
The study was approved by the IEC for Clinical Research of Zhongda Hospital, Affiliated to Southeast University (Approval Number: 2022ZDSYLL218-P01). Informed consent was obtained from all subjects and/or their legal guardian(s).

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests.

Received: 25 March 2024 / Accepted: 8 July 2025
Published online: 01 September 2025

References

1. Rubioochoa J, Bentezmartnez J, Lluch E, et al. Physical examination tests for screening and diagnosis of cervicogenic headache: A systematic review[J]. Man Therap. 2016;21:35–40. https://doi.org/10.1016/j.math.2015.09.008.
2. Wang XL, Liu J, Li ZQ, et al. Application of Physical Examination Data on Health Analysis and Intelligent Diagnosis[J] Biomed Res Int. 2021;8828677. https://doi.org/10.1155/2021/8828677
3. Latanya Sweeney. K-anonymity: a model for protecting privacy[J]. Int J Uncertain Fuzziness Knowledge-Based Syst. 2002;10(05):557–70. https://doi.org/10.1142/S0218488502001648.
4. Kumar P, Mayil Vel M, Karthikeyan. L-diversity on k-anonymity with external database for improving privacy preserving data publishing[J]. Int J Comput Appl. 2012;54(14):7–13. https://doi.org/10.5120/8632-2341.
5. Machanavajjhala A, Kifer D, et al. L-diversity: privacy beyond k-anonymity[J]. Acm Trans Knowl Discovery Data. 2007;1(1):3. https://doi.org/10.1145/1217299.1217302.
6. Wang Qian Z, Xu S, Qu. An enhanced k-anonymity model against homogeneity attack[J]. J Softw. 2011;6(10):1945–52. https://doi.org/10.4304/JSW.6.10.1945-1952.
7. Li N, Li T. Suresh Venkatasubramanian. t-Closeness: privacy beyond k-anonymity and l-diversity[J]. International Conference on Data Engineering IEEE, 2007:106–115. https://doi.org/10.1109/ICDE.2007.367856
8. Riyana S, Nanthachumphu S, Riyana N. Achieving privacy preservation constraints in Missing-Value Datasets[J]. SN COMPUT SCI. 2020;1(4). https://doi.org/10.1007/s42979-020-00241-9.
9. Xiao, Xiaokui and Yufei Tao. Anatomy: simple and effective privacy preservation [C]. In Proceedings of the 32nd international conference on Very large data bases (VLDB 2006). Seoul: 2006:139–150.
10. Riyana S. Achieving anatomization constraints in dynamic Datasets[J]. ECTI transactions on computer and information technology (ECTI-CIT), 2023, 17(1): 27–45. https://doi.org/10.37936/ecti-cit.2023171.248877
11. Riyana S, (l^ p_1. l^ p_n Lp 1,…, Lpn)-Privacy: privacy preservation models for numerical quasi-identifiers and multiple sensitive attributes[J]. J Ambient Intell Humaniz Comput. 2021;12:9713–29. https://doi.org/10.1007/s12652-020-02715-3.
12. Riyana S, Sasujit K, Homdoung N. ECTI Trans Comput Inform Technol (ECTI-CIT). 2023;17(3):452–68. https://doi.org/10.37936/ecti-cit.2023173.252952. Privacy-enhancing data aggregation for big data analytics[J].
13. Zhang Q, Koudas N, Srivastava D et al. Aggregate Query Answering on Anonymized Tables[C]. 2007 IEEE 23rd International Conference on Data Engineering. Istanbul, Turkey, 2007: 116–125. https://doi.org/10.1109/ICDE.2007.367857
14. Riyana S, Riyana NE. SN Comput sci. 2024;5:340. https://doi.org/10.1007/s42979-024-02690-y. Security and Privacy Preservation for the Publication of Rating Datasets[J].
15. Dwork C. Differential privacy[C]. Proc of the 33rd int colloquium on automata, languages and programming (ICALP 2006). Berlin: Springer,2006:1–12. https://doi.org/10.1007/11787006_1
16. Dwork C, McSherry F, Nissim K, et al. Calibrating noise to sensitivity in private data analysis[C]. Proc of the 3rd theory of cryptography Conf (TCC 2006). Berlin: Springer; 2006. pp. 363–85. https://doi.org/10.1007/11681878_14.
17. Singh AK, Gupta R. A privacy-preserving model based on differential approach for sensitive data in cloud environment[J]. Multimed Tools Appl. 2022;81(23):33127–50. https://doi.org/10.1007/s11042-021-11751-w.
18. Gupta R, Saxena D, Gupta I, Singh AK. IEEE Netw Lett. 2022;4(4):217–21. https://doi.org/10.1109/LNET.2022.3215248. Differential and TriPhase Adaptive Learning-Based Privacy-Preserving Model for Medical Data in Cloud Environment[J].
19. Gupta R, Singh AK. New Gener Comput. 2022;40(3):737–64. https://doi.org/10.1007/s00354-022-00185-z. A Differential Approach for Data and Classification Service-Based Privacy-Preserving Machine Learning Model in Cloud Environment[J].
20. Li W, Zhang X, Li X, et al. PPDP-PCAO: an efficient high-dimensional data releasing method with differential privacy protection[J]. IEEE Access. 2019;7:176429–37. https://doi.org/10.1109/ACCESS.2019.2957858.
21. Xu C, Ren J, Zhang Y, et al. DPPro: differentially private high-dimensional data release via random projection[J]. IEEE Trans Inf Forensics Secur. 2017;12(12):3081–93. https://doi.org/10.1109/TIFS.2017.2737966.
22. Jinxin H, Yingjie W, Jianping C, et al. J Comput Res Dev. 2022;59(01):182–96. https://doi.org/10.7544/issn1000-1239.20200701. Differentially private high-dimensional binary data publication via attribute segmentation [J].
23. Kumar G, Sathish, et al. No more privacy concern: A privacy-chain based homomorphic encryption scheme and statistical method for privacy preservation of user's private and sensitive data[J]. Expert Syst Appl. 2023;234:121071. https://doi.org/10.1016/j.eswa.2023.121071.
24. Kumar G, Sathish, et al. Differential privacy scheme using Laplace mechanism and statistical method computation in deep neural network for privacy preservation[J]. Eng Appl Artif Intell. 2024;128:10739. https://doi.org/10.1016/j.engappai.2023.107399.
25. Kumar GS, Premalatha K. STIF: intuitionistic fuzzy Gaussian membership function with statistical transformation weight of evidence and information value for private information preservation[J]. Distrib Parallel Databases. 2023;1–34. https://doi.org/10.1007/s10619-023-07423-3.
26. Kumar G, Sathish, Kandhasamy Premalatha. Securing private information by data perturbation using statistical transformation with three dimensional shearing[J]. Appl Soft Comput. 2021;112:107819. https://doi.org/10.1016/j.asoc.2021.107819.
27. Jun ZHANG, CECILIA GRAHAMC. ACM Trans DATABASE Syst. 2017;42(4):1–41. https://doi.org/10.1145/3134428. M, et al. PrivBayes: Private Data Release via Bayesian Networks[J].
28. Biao X, Hongqiang Y, Haining L, et al. Research on improvement of bayesian network privacy protection algorithm based on differential privacy[J]. Netinfo Secur. 2020;20(11):75–86. https://doi.org/10.3969/j.issn.1671-1122.2020.11.010.
29. Park Y, Joydeep Ghosh. PeGS: Perturbed Gibbs Samplers that Generate Privacy-Compliant Synthetic Data [J]. ACM. 2014;7(3):253–82. https://doi.org/10.14778/3407790.3407802.

30. Judea P. Bayesian networks: A model of self-activated memory for evidential reasoning[C]. Proceedings of the 7th Conference of the Cognitive Science Society, 1985: 329–334.
31. Heikkilä M, Jälkö J, Dikmen O, et al. Differentially private Markov chain Monte Carlo[J]. Adv Neural Inf Process Syst. 2019;32:4115–25. https://doi.org/10.48550/arXiv.1901.10275.
32. Rodrigues GS, Nott DJ, Sisson SA. Likelihood-free approximate Gibbs sampling[J]. Stat Comput. 2020;30(4):1057–73. https://doi.org/10.1007/s11222-020-09933-x.
33. Pai MM, Roth A. Privacy and mechanism design[J]. ACM SIGecom Exchanges. 2013;12(1):8–29. https://doi.org/10.1145/2509013.2509016.
34. Mcsherry FD. Privacy integrated queries: An extensible platform for privacy-preserving data analysis[C]. 2009 ACM SIGMOD International Conference on Management of data, 2009: 19–30. https://doi.org/10.1145/1559845.1559850
35. Mcsherry F, Talwar K. Mechanism design via differential privacy[C]. Proceedings of the 48th Annual IEEE Symposium on Foundations of Computer Science, 2007: 94–103. https://doi.org/10.1109/FOCS.2007.41
36. Weinberger K, Dasgupta A, Langford J et al. Feature hashing for large scale multitask learning[C]. In Proceedings of the 26th International Conference on Machine Learning, 2009:1113–1120. https://doi.org/10.1145/1553374.1553516
37. McKenna R, Sheldon D, Miklau G. Winning the NIST contest: A scalable and general approach to differentially private synthetic data[J]. J Priv Confidentiality. 2021;11(3). https://doi.org/10.29012/jpc.778.
38. Hershey JR, Olsen PA. Approximating the Kullback Leibler Divergence Between Gaussian Mixture Models[J]. 2007 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 07). Honolulu, HI, USA, 2007, 4: 317–320. https://doi.org/10.1109/ICASSP.2007.366913
39. Lionel Blanchet R, Vitale R, van Vorstenbosch et al. Constructing bi-plots for random forest: Tutorial[J]. Anal Chim Acta 2020,1131:146–55. https://doi.org/10.1016/j.aca.2020.06.043
40. Jiang D, Zhang G, Karami M et al. DP2-VAE: differentially private Pre-trained variational autoencoders [DB/OL]. (2022-11-2) [2025-5-1]. ArXiv, abs/2208.03409.
41. Stella Ho Y, Qu B, Gu, et al. DP-GAN: differentially private consecutive data publishing using generative adversarial nets[J]. J Netw Comput Appl. 2021;185. https://doi.org/10.1016/j.jnca.2021.103066.
42. Xiyang Liu W, Kong P, Jain et al. DP-PCA: Statistically Optimal and Differentially Private PCA [DB/OL]. (2022-5-27) [2025-5-1]. ArXiv, abs/2205.13709.
43. Swope R, Khanna A, Doldo P et al. Feature Selection from Differentially Private Correlations [DB/OL]. (2024-8-23) [2025-5-1]. ArXiv, abs/2408.10862.

Publisher's note