Annals of Data Science
https://doi.org/10.1007/s40745-023-00475-3

A Survey on Differential Privacy for Medical Data Analysis

WeiKang Liu[1] · Yanchun Zhang[1,2] · Hong Yang[1] · Qinxue Meng[3]

Received: 28 February 2023 / Revised: 16 May 2023 / Accepted: 22 May 2023

Abstract

Machine learning methods promote the sustainable development of wise information technology of medicine (WITMED), and a variety of medical data brings high value and convenience to medical analysis. However, the applications of medical data have also been confronted with the risk of privacy leakage that is hard to avoid, especially when conducting correlation analysis or data sharing among multiple institutions. Data security and privacy preservation have recently played an essential role in the field of secure and private medical data analysis, where many differential privacy strategies are applied to medical data publishing and mining. In this paper, we survey research work on the applications of differential privacy for medical data analysis, discussing the necessity of medical privacy-preserving, the advantages of differential privacy, and their applications to typical medical data, such as genomic data and wearable device data. Furthermore, we discuss the challenges and potential future research directions for differential privacy in medical applications.

Keywords Privacy computing · Differential privacy · Medical data · Data publishing

✉ WeiKang Liu
dpstudier@e.gzhu.edu.cn

Yanchun Zhang
yanchun.zhang@vu.edu.au

Hong Yang
hyang@gzhu.edu.cn

Qinxue Meng
Qinxue.meng@gmail.com

[1] Cyberspace Institute of Advanced Technology, Guangzhou University, Guangzhou, China

[2] Institute of Sustainable Industries and Liveable Cities, Victoria University, Melbourne, Australia

[3] College of Information Engineering, Suzhou University, Suzhou, China

Published online: 10 June 2023

1 Introduction

Recently, we have witnessed an increasing number of data science applications in sustainable development field of wise information technology of medicine (WITMED). More applications include drug discovery and disease surveillance, where personal information such as name, age, gender, postal code, profession, disease, and medical history can be collected, published and used by third-party terminal devices or authorities. The analysis and applications of medical data have become a hot topic in recent years [1, 2].

Combining data science and modern medicine, the benefits of analyzing medical data span disease prediction, new drug research and development, auxiliary diagnosis and treatment, and health management. However, as more data is collected and processed through interconnected devices [3], privacy becomes a significant concern due to private sensitive information that may be contained within the data.

In data science research, data privacy-preserving has become increasingly significant in addressing security and privacy challenges. The development of privacy-enhancing techniques, including differential privacy, secure multi-party computation and homomorphic encryption, is imperative for enabling privacy protection while collecting and analyzing data collaboratively. Additionally, transparent and accountable data governance frameworks that protect privacy and facilitate informed consent should be developed to ensure the responsible utilization of data. Therefore, adopting a comprehensive approach that encompasses both technical and ethical considerations is necessary to effectively address the privacy challenges that arise at the intersection of artificial intelligence and data science.

As for medical data analysis, we have observed that the attacks on medical datasets and models have increased rapidly in recent years. Therefore, the research on privacy-preserving methods has become a crucial area of study in medical informatics field. Privacy computing can realize medical simulation, prediction and security statistical analysis of medical data with specific privacy-preserving levels. For publishing medical data, anonymous methods are capable of defending against linking attacks, skewness attacks and similarity attacks, to name a few. However, they do not have enough resistance to background knowledge [4]. Differential privacy is not only robust to differential attacks, but also defending against all of the above attacks on medical sensitive data. Moreover, for publicly published models, differential privacy algorithms also prevent adversarial recovery of private information from the original medical data.

In recent years, there has been a surge in development of novel algorithms for differential privacy medical analysis, which this paper aims to conduct a survey on. And the efforts of this paper can be summarized as follows. First, we discuss why differential privacy is considerable in medical data publishing data and data mining. Second, we discuss typical differential privacy methods based on noises, which can help better understand existing work. Third, we analyze the limitations of differential privacy strategy and summarize possible future challenges, highlighting future research directions of medical applications of differential privacy.

The rest of this paper is structured as follows. Section 2 introduces privacy computing technology of medical data and the characteristics of anonymous methods. The fundamental theories of differential privacy and its noise mechanisms are obtained in Sect. 3. Section 4 illustrates applications of differential privacy to medical data. Subsequently, we analyze and discuss partial possible future challenges of differential privacy in Sect. 5. Section 6 concludes the paper.

2 Medical Data Privacy Computing

The development of privacy connotation is dynamic, continuing to enrich its meaning with the progress of social politics, economic culture and the improvement of human consciousness. The so-called privacy computing is a series of privacy-preserving methods that protect sensitive data from being visible but available when using conjoint analysis and computing collaboratively on model data.

Unlike secure blockchain framework [5] or some web attack detection techniques in cloud-IoT system [6], privacy computing has mainly integrated cryptography, artificial intelligence and computer hardware technologies into a relatively mature technical system represented by multi-party security computation, trusted execution environment and federated learning. Meanwhile, it also regards differential privacy, homomorphic encryption, zero-knowledge proof and others as auxiliary technology, providing a technical guarantee for data security and circulation.

The research on privacy problems can be divided into five categories: financial privacy, Internet privacy, medical privacy, political privacy and information privacy [7]. Among them, medical privacy comes from a wide range of sources and complex types of medical data, mainly including information that patients do not want to be known to the outside world, such as genomic information, past medical history, medical records, etc. They are commonly stored in the form of electronic medical record (EMR), electronic health records (EHR) and personal health records (PHR).

Medical data scattered in different institutions is difficult to interconnect each part, which may seriously restrict the output of clinical scientific research results. For this problem, privacy computing technology has the ability to provide a series of practical solutions to achieve data circulation and take full advantage of medical data. What's more, it can also solve the problem of insufficient samples from a single institution that leads to credibility loss of research results.

During the COVID-19 epidemic prevention and control period [8, 9], analyses on medical services and tests, pulse count, body temperature and the overall effect of age and gender was done [10, 11]. Furthermore, the use of privacy computing technology such as multi-party security computing enables researchers from all over the world to jointly conduct genome analysis of case samples and share sequencing results without disclosing detailed personal information, so as to implement real-time tracking of the current virus situation and prediction of future strain evolution [1, 12]. This will help more countries diagnose COVID-19 patients efficiently and take effective measures in time.

Generally, genome analysis relies on a large number of personally private data. Using privacy computing will have the original genetic data sealing in local database and realize safe sharing of sensitive genomic data. Then the joint calculation and

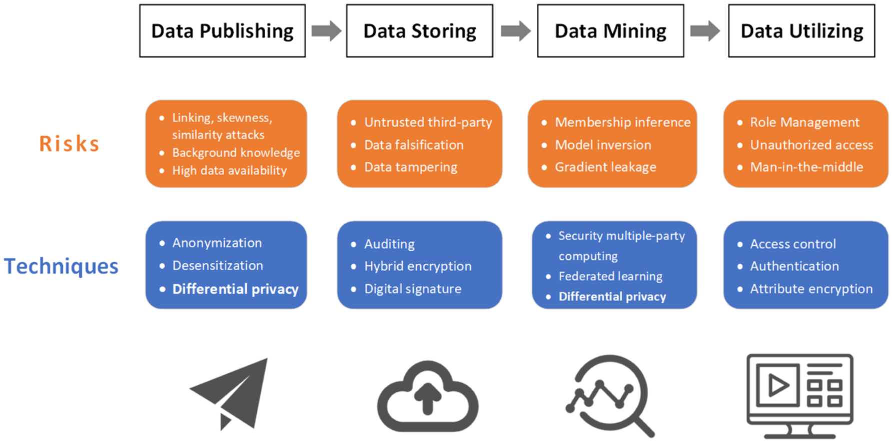

Fig. 1 Privacy-preserving life cycle of medical data

association analysis will be carried out. In this way, various genome resources can be mined by different medical institutions under the premise of privacy-preserving.

For clinical medical research, utilizing local data protected by privacy computing technology can implement distributed statistical analysis algorithms to joint modeling and obtain related results, such as feasibility analysis of clinical research, cohort study with large samples, disease prediction and drug insight, etc. Therefore, the application of privacy computing will greatly improve medical research efficiency and accelerate the transformation of scientific research achievements.

As shown in Fig. 1, the complete medical data life cycle incorporates data publishing, storing, mining and utilizing [13, 14]. Data publishers, storage parties, miners and users are involved in this process. Both the private data threats and the corresponding privacy-preserving techniques are different at each phase.

In practical medical scenarios, the data publishing phase usually involves continuous release of medical data, and attracts the attention of adversaries, who are able to combine specific background knowledge to carry out a series of analyses and attacks on sensitive medical data. Thus, in data publishing phase, while ensuring efficient transmission and strong usability of data, considering how to safely and reliably deal with sensitive information which may be leaked is also a crucial issue for medical researchers and clinicians.

Traditional anonymous publishing methods are usually adopted in the process of medical data release, including k-anonymity [15], l-diversity [16] and t-closeness [17]. Through generalization, suppression and substitution of dataset tuples, they align identifier classification based on specific rules so as to meet the need for medical data desensitization release. Although anonymous approaches are capable of protecting sensitive plaintext information [18, 19], they cannot effectively prevent the attackers from using background knowledge depending on external databases to link attacks, and their privacy protection effect lacks strict theoretical proof. Exactly, the differential privacy computing technology mainly introduced in the following can make up for the disadvantages of anonymization methods to solve corresponding problem.

3 Differential Privacy

In a hypothetical scenario, if data collectors have to collect the published patient diagnosis and treatment records from a hospital, differential privacy can protect sensitive information by adding random noise or disturbance to the original records, which not only cannot reveal certain personal data of a certain user in the datasets, but also ensures the overall statistical characteristics within specified bounds, thus maintaining data utility to a certain extent. That strategy greatly ensures the privacy and security of medical data.

Proposed by Dwork et al. [20], the concept of differential privacy comes from semantic security in cryptography. On the one hand, differential privacy makes it impossible for adversaries to distinguish the encryption results of different plaintexts. On the other hand, it provides a strict upper limit of privacy protection in mathematics, that is, privacy budget. To prevent differential attacks by adding random noise is the direct purpose of differential privacy, so that the adversary cannot effectively infer personal privacy while maximizing the availability of query results in neighboring datasets. The differential attack is that the adversary makes use of subtraction thinking in neighboring datasets to infer sensitive data of a certain person by comparing statistical results of queries.

In the data publishing phase, using differential privacy can ensure that one same data is queried in two neighboring datasets and the results are basically the same, so as to confuse the judgment of the adversary. In addition to guarding against differential attacks, differential privacy can also prevent link attacks based on background knowledge to a large extent.

3.1 Definition

Generally speaking, differential privacy is defined as follows: Given a randomized algorithm (query function) M, P_m is the set of all range values that M outputs, and $S_m \subseteq P_m$. For any two neighboring datasets D and D' (at most differing on one-row data), if the algorithm M satisfies:

$$\Pr[M(D) \in S_m] \leq e^{\varepsilon} \cdot \Pr\left[M\left(D'\right) \in S_m\right] \tag{3.1}$$

Then it is said that algorithm M satisfies ε-differential privacy, where the parameter ε is the privacy budget. As can be seen from Eq. (3.1) (or put e^{ε} on the right side alone), the smaller the privacy budget is, the probability distribution of query results returned by M on neighboring datasets is more similar, accompanied by the harder it is for the adversary to distinguish the pair of neighboring datasets. It provides higher protection degree of sensitive data, but correspondingly, data utility will get worse gradually. On the contrary, a larger privacy budget will lower the degree of privacy protection and improve data utility.

Notably, the probabilities of the third party querying neighboring datasets to get the same statistic value are only very close, not exactly equal. While protecting specific data from leakage, it is also essential to prevent the data from being completely randomized, leading to the loss of usability.

3.2 Noise-Based Mechanisms

In this part, we discuss three noise mechanisms commonly used in differential privacy.

3.2.1 Laplace Mechanism

The query request of the original dataset D is regarded as the value of a function f on D. Laplace mechanism is achieved by adding noise η to $f(D)$ and the result is $f(D) + \eta$. η is a continuous random variable satisfying $Lap\left(0, \frac{\Delta(f)}{\varepsilon}\right)$ distribution and its probability density function is:

$$P(\eta) = \frac{1}{2\lambda}e^{-\frac{|\eta|}{\lambda}} \tag{3.2}$$

In Eq. (3.2), the expected value of the Laplace distribution is 0, the variance is $2\lambda^2$, and the parameter λ reflects the amplitude of noise and the intensity of privacy protection. Larger λ means the greater range of noise added and the higher degree of privacy protection. In addition, the sensitivity is also an important factor affecting the strength of privacy protection.

Given a query function f, if $f : D \to R$ (query result), the global sensitivity of f is:

$$\Delta(f) = \max_{D,D'} \| f(D) - f\left(D'\right) \|_1 \tag{3.3}$$

for all neighboring datasets D and D*'*.

The global sensitivity reflects the maximum range of variation of a query function over neighboring datasets, in conjunction with privacy budgets to control the amount of generated noise.

3.2.2 Gaussian Mechanism

The Gaussian noise is a mechanism to achieve (ε, δ)-differential privacy, which is defined as follows:

$$\Pr[M(D) \in S_m] \le e^{\varepsilon} \cdot \Pr\left[M\left(D'\right) \in S_m\right] + \delta \tag{3.4}$$

Here in (3.4), the additive term δ denotes the probability of violating plain ε-differential privacy is allowed. Given a function f over dataset D, if $\varepsilon < 1, \delta \in (0, 1)$ and $\delta \ge \frac{4}{5}e^{\frac{-(\sigma\varepsilon)^2}{2}}$ [21], $\delta > \sqrt{2ln\frac{1.25}{\delta}}\Delta f/\varepsilon$, Gaussian noise mechanism can be expressed as: $M(D) = f(D) + N\left(0, \Delta f^2 \cdot \sigma^2\right)$ [22, 23], N is the standard Gaussian distribution with zero-mean Gaussian noise parameter σ and a standard deviation of $\Delta f \cdot \sigma$. Compared with L_1-sensitivity norm used by Laplace mechanism, Gaussian mechanism follows the same privacy composition, but uses the L_2-sensitivity norm.

3.2.3 Exponential Mechanism

The above two noise mechanisms are mainly used to protect numerical data, while the exponential mechanism is suitable for non-numerical data. It defines a practical evaluation function q, in charge of calculating a satisfaction score ω for each output scheme. The scheme with high score will have a higher probability to be published, the exponential mechanism satisfies:

$$\Pr(\omega) \propto \exp\left(\frac{\varepsilon}{2\Delta(q)} q(D, \omega)\right) \tag{3.5}$$

In Formula (3.5), $\Delta(q)$ is the global sensitivity of the evaluation function.

3.3 Classification of Differential Privacy

Traditional differential privacy will gather the original datasets to a data center and then release relevant statistical information satisfying differential privacy, which is called centralized differential privacy (CDP). In other words, CDP's protection of sensitive information has always been based on the assumption that the third-party data collectors are trusted, that is, they will not steal or disclose sensitive information from users. However, in practical applications, users' privacy is still not guaranteed [24]. An investigation in 2018 showed that most mobile health apps jeopardized users' privacy by violating data protection regulations and revealing sensitive information [25].

In view of this, local differential privacy (LDP) [26] emerges in the scenario of untrusted third-party data collectors. When suffering the same quantified privacy attacks of CDP, LDP will subdivide the protection of sensitive personal information. Specifically, LDP delivers data protection authority to each user, enabling users to protect sensitive personal information independently, thus achieving more thorough privacy preservation locally. At present, LDP has been mainly used in frequency estimation, mean estimation [27] and gradually been put into industrial applications. For example, Apple [28] applied it in iOS 10 operating system to protect user device data, and Google [29] used it to collect users' behavior statistics from the Chrome browser.

3.4 Differential Privacy in Machine Learning

Recently, differential privacy has also been gradually applied in data mining field and combined with increasing machine learning algorithms.

Differential privacy depends on noise or disturbance, so compared with other privacy computing methods, it has low computational complexity, improving its application efficiency in the field of machine learning while providing more explicit privacy guarantees. Noises can not only be added to original data, objective function, output model parameters or features extracted by neural network [30], but also be disturbed or screened for sensitive features specified by users or automatically detected by the recognition network [31, 32]. Shokri et al. [33] used differential privacy mechanisms to design a distributed learning method for privacy protection early on. In their method, privacy loss can be calculated according to the parameters of the model, but too many model parameters may lead to huge privacy loss. On this basis, Abadi et al. [22] improved it and introduced a more efficient gradient descent algorithm based on differential privacy, which has a smaller privacy budget and better performance. More importantly, Abadi et al. [22] also introduced a measuring method of privacy loss, Moment Accountant, to automate the calculation of privacy loss. The differential privacy stochastic gradient descent (DP-SGD) algorithm mentioned in the paper also laid the foundation for more scholars researching on machine learning of privacy protection in the future.

Applying differential privacy to machine learning will reduce the probability that the adversary can reversely deduce sensitive personal information from the model in the original training datasets. Data utility and model security are both crucial in this process. On the one hand, it is necessary to reasonably select and control the privacy budget in the training process according to the privacy loss. Methods such as dynamic allocation of privacy budget [34], utilizing differential privacy post-processing property for noise reduction [35], or reducing privacy budget that may be caused by combination characteristics [36] can be considered. On the other hand, some model architectures that are more conducive to protecting user privacy can also be selected [37, 38].

4 Differential Privacy for Medical Data

In medical data, differential privacy is mainly applied to data publishing and data mining. In the data publishing phase, it can greatly prevent the privacy leakage caused by the data query based on background knowledge. In the data mining phase, it can resist the privacy leakage caused by the membership inference attack (MIA) of the adversary on the model.

As Fig. 2 shows, current applications research focuses on genomic data, medical wearable devices, electronic medical records and medical images, etc.

4.1 Genomic Data

Genomic data in medicine is DNA sequence with genetic benefits of individuals, such particular data is difficult to change over the life cycle and of long-lived value [39–41]. Given this, some enterprises may be tempted by commercial interests to violate the genetic privacy of others.

Genome-wide association study (GWAS) is conducive to learning genome-phenome associations by analyzing the statistical correlation between the variants of a case group (phenotype positive) and a control group (phenotype negative) [4]. The adversary may infer the potential traits and genotypes of victims depending on trait associations available from GWAS catalogue [42]. In order to reduce the possibility

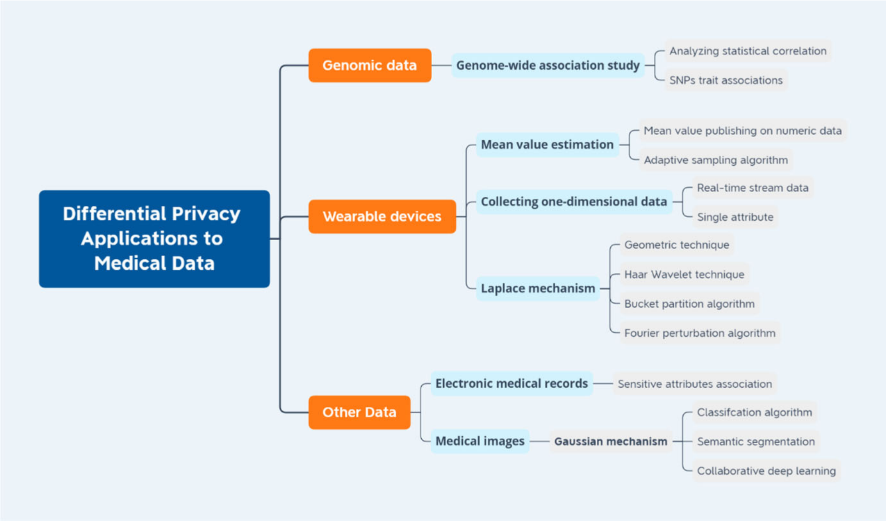
Differential Privacy Applications to Medical Data
Genomic data
Genome-wide association study
Analyzing statistical correlation
SNPs trait associations
Wearable devices
Mean value estimation
Mean value publishing on numeric data
Adaptive sampling algorithm
Collecting one-dimensional data
Real-time stream data
Single attribute
Laplace mechanism
Geometric technique
Haar Wavelet technique
Bucket partition algorithm
Fourier perturbation algorithm
Other Data
Electronic medical records
Sensitive attributes association
Medical images
Gaussian mechanism
Classifcation algorithm
Semantic segmentation
Collaborative deep learning

Fig. 2 Differential privacy application to medical data

of leaking genome privacy from published aggregate statistics of GWAS, differential privacy strategies can be widely introduced in it. For example, to a certain extent, differential privacy can prevent attackers from inferring the number and location of single nucleotide polymorphisms (SNPs) that might be significantly linked with certain diseases in the original genetic datasets, so as to protect the gene privacy [43, 44]. For another example, the controlled noise in differential privacy can be added to query results from genomic database, which promotes genome openness while preserving privacy [45, 46]. However, large scale of added noise to high-dimensional genomic data will inevitably degrade data utility. To address this problem, He et al. [47] proposed an effective method to factorize a huge-dimensional distribution into a set of local distributions, reducing the scale of added noise.

Moreover, Almadhoun et al. [48] showed that the adversary could infer genome privacy from query results added noise by exploiting the correlations between the genomes of family members with dependency, then Almadhoun et al. [49] formalized the differential privacy notion to avoid sensitive information inference by adversary relying on tuples prior knowledge. Similar to this work, in order to strengthen the effect of differential privacy against correlation attacks, Yilmaz et al. [50] proposed a scheme which eliminates certain states of a SNP loosely correlated with previously shared SNPs. Chen et al. [51] researched on machine learning model’s ability to defend against MIA on genomic data and evaluated the effect of model sparsity on privacy vulnerability with different differential privacy settings.

4.2 Wearable Device Data

Medical wearable devices storing personal health data such as heart rate and blood sugar play an important role in disease diagnosis and treatment, and they made it possible to collect real-time medical health data continuously [52]. Personal sensitive data stored in medical wearable devices need to be collected in real time, they also have a demand for privacy-preserving in data publishing.

Tu et al. [53] applied differential privacy to numerical mean stream data publishing of medical wearable devices, and adopted an adaptive sampling algorithm based on Kalman filter adjustment error to allocate privacy budgets, which improves the usability of published stream data. Kim et al. [54] added Laplace noise to salient points for collecting one-dimensional heart rate data, but existing large data error.

Revolving around Laplace mechanism, researchers have extended a series of works to provide better data utility and privacy guarantee. Li et al. [55] proposed an improved randomized method to tackle stream medical data collection with a single attribute. That method incorporates random response and Laplace mechanism, further improving the availability of mean value estimation with stream data in medical wearable devices. Moreover, for partitioning or temporal medical datasets, the geometric technique [56], Haar Wavelet technique [57], bucket partition algorithm [58] and Fourier perturbation algorithm [59] have also been adopted to combine with Laplace distribution of differential privacy.

4.3 Other Medical Data

As an inevitable product of modern information technology in the medical field, the electronic medical record is the carrier of various medical information in diagnosis and treatment process, greatly benefiting modern management of hospital medical records. Combining with LDP strategy, Wu et al. [60] designed a blockchain-enabled framework to provide attribute-based privacy protection for transactions. Medical diagnosis results also belong to a part of electronic medical records, Chen et al. [61] presented a differential privacy quasi-identifier classification scheme to tackle original disease dataset and defined privacy ratio for evaluating dataset vulnerability. Zhang et al. [62] designed an attribute association-based differential privacy classification tree method of data publishing, conducting experiments on real medical record datasets.

In addition, Ziller et al. [63] proposed an open-source software framework based on DP-SGD algorithm application to deal with medical imaging classification and semantic segmentation deep learning tasks. Yuan et al. [64] exploited collaborative deep learning with Gaussian noise mechanism to experiment on X-ray Images (Pneumonia) dataset and found the accuracy loss was small, affecting little to the results. Adnan et al. [65] indicated that federated learning with differential privacy has been the viable and reliable collaborative machine learning framework for medical image analysis.

5 Discussions

Although differential privacy to medical data has made some achievements at present, it still faces difficulties and challenges in terms of practical application.

Firstly, we still need to explore how to constantly improve data utility when medical data is shared and circulated across institutions, and to select suitable algorithm strategies to reduce global sensitivity and control privacy budget.

Second, due to the complexity of the scale and structure to medical data, rapidly increasing medical data has begun to be expressed in an unstructured form. As a popular method to describe networked data [66, 67], graph neural network (GNN) has also been successively applied to kinds of medical tasks by plenty of researchers, such as predicting chemical properties of molecules, biological interaction properties of proteins, drug recommendation, etc. [68–70]. However, when the GNN models are uploaded to the server and the graph nodes or labels involve personal sensitive information, the process of learning graph data still has the possibility of privacy leakage. For this scenario, differential privacy strategy can also be used to add noise locally [71, 72]. Combined with differential privacy, graph data has a more complex structure than general medical data types. On the one hand, the structural characteristics of the graph may extremely increase the global sensitivity of queries, resulting in excessive noises. On the other hand, since each user locally perturbs the data independently, how to ensure the relevance between original data and then build a graph structure with high availability based on disturbed data also become the main challenges in current practical applications.

Third, existing privacy-preserving computation methods have their own limitations. Finding a reasonable trade-off between privacy-preserving intensity, data utility and algorithm execution efficiency has always been the common goal of these methods [73–75]. Regarding differential privacy as a privacy-enhancing technique to combine with mainstream privacy computing methods like federated learning can be considered and widely applied to distributed training of decentralized medical data in the future.

6 Conclusion

Due to increasingly large scale and complex structure, medical data contains sensitive personal information inevitably, and the demand for privacy-preserving is particularly prominent. In this survey, we discussed the development of differential privacy and its applications to medical data. As a privacy computing method with strict mathematical limitations and various implementations, differential privacy is capable of solving the security and efficiency challenges during medical data publishing and mining. Moreover, this work provided a reliable environment and solution for medical data analysis. Finally, we discussed major challenges and future research directions about the medical data applications of differential privacy.

Author Contributions The first author conducted survey and wrote the manuscript. The second author and the third author reviewed the work and edited the manuscript. All authors read and approved the final manuscript.

Funding This work was supported by the Joint Research Fund of Guangzhou and University under Grants Nos. 202201020218 and 202201020380 and Guangdong Higher Education Innovation Group 2020KCXTD007.

Data Availability Data sharing is not applicable to this article as no datasets were generated or analyzed during the current study.

Code Availability Code sharing is not applicable to this article as no new code was created or used.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

Ethical Approval All of the followed procedures were in accordance with the ethical and scientific standards. This article does not contain any studies with human participants performed by the author.

References

1. Belle A, Thiagarajan R, Soroushmehr SM et al (2015) Big data analytics in healthcare. BioMed Res Int. https://doi.org/10.1155/2015/370194
2. Shi Y (2022) Advances in big data analytics: theory, algorithms and practices. Springer, Singapore
3. Tien JM (2017) Internet of things, real-time decision making, and artificial intelligence. Ann Data Sci 4(2):149–178. https://doi.org/10.1007/s40745-017-0112-5
4. Sun Z, Wang Y, Shu M et al (2019) Differential privacy for data and model publishing of medical data. IEEE Access 7:152103–152114. https://doi.org/10.1109/ACCESS.2019.2947295
5. Tian Z, Li M, Qiu M et al (2019) Block-DEF: a secure digital evidence framework using blockchain. Inf Sci 491:151–165. https://doi.org/10.1016/j.ins.2019.04.011
6. Tian Z, Luo C, Qiu J et al (2019) A distributed deep learning system for web attack detection on edge devices. IEEE Trans Ind Inform 16(3):1963–1971. https://doi.org/10.1109/TII.2019.2938778
7. Fang B, Jia Y, Li A et al (2016) Privacy preservation in big data: a survey. Big Data Res 2(1):1–18. https://doi.org/10.11959/j.issn.2096-0271.2016001
8. Li J, Guo K, Herrera Viedma E, Lee H, Liu J, Zhong Z, Gomes L, Filip FG, Fang SC, Özdemir MS, Liu XH, Lu G, Shi Y (2020) Culture vs policy: more global collaboration to effectively combat COVID-19. The Innovation 1(2):100023. https://doi.org/10.1016/j.xinn.2020.100023
9. Liu Y, Gu Z, Xia S, Shi B, Zhou X, Shi Y, Liu J (2020) What are the underlying transmission patterns of COVID-19 outbreak? An age-specific social contact characterization. EClinicalMedicine 22:100354. https://doi.org/10.1016/j.eclinm.2020.100354
10. Radanliev P, De Roure D, Walton R et al (2022) What country, university, or research institute, performed the best on Covid-19 during the first wave of the pandemic? Ann Data Sci 9(5):1049–1067. https://doi.org/10.1007/s40745-022-00406-8
11. Gada V, Shegaonkar M, Inamdar M et al (2022) Data analysis of COVID-19 hospital records using contextual patient classification system. Ann Data Sci 9(5):945–965. https://doi.org/10.1007/s40745-022-00378-9
12. Yan S, Lv A (2021) Overview of the development of privacy preserving computing. Inf Commun Technol Policy 47(6):1–11. https://doi.org/10.12267/j.issn.2096-5931.2021.06.001
13. Olson DL, Shi Y (2007) Introduction to business data mining. McGraw-Hill/Irwin, New York
14. Shi Y, Tian YJ, Kou G, Peng Y, Li JP (2011) Optimization based data mining: theory and applications. Springer, Berlin
15. Sweeney L (2002) k-anonymity: a model for protecting privacy. Int J Uncertain Fuzziness Knowl Based Syst 10(05):557–570. https://doi.org/10.1142/S0218488502001648
16. Machanavajjhala A, Kifer D, Gehrke J, Venkitasubramaniam M (2007) l-diversity: privacy beyond k-anonymity. ACM Trans Knowl Discov Data 1(1):3-es. https://doi.org/10.1145/1217299.1217302
17. Li N, Li T, Venkatasubramanian S (2007) t-closeness: privacy beyond k-anonymity and l-diversity. In: 2007 IEEE 23rd international conference on data engineering. IEEE, pp 106–115. https://doi.org/10.1109/ICDE.2007.367856
18. Ge YF, Wang H, Cao J et al (2022) An information-driven genetic algorithm for privacy-preserving data publishing. In: Web information systems engineering–WISE 2022: 340–354. https://doi.org/10.1007/978-3-031-20891-1_24

19. Ge YF, Zhan ZH, Cao J et al (2022) DSGA: a distributed segment-based genetic algorithm for multi-objective outsourced database partitioning. Inf Sci 612:864–886. https://doi.org/10.1016/j.ins.2022.09.003
20. Dwork C (2006) Differential privacy. In: ICALP 2006: automata, languages and programming, pp 1–12. https://doi.org/10.1007/11787006_1
21. Dwork C, Roth A (2014) The algorithmic foundations of differential privacy. Found Trends Theor Comput Sci 9(3–4):211–407. https://doi.org/10.1561/0400000042
22. Abadi M, Chu A et al (2016) Deep learning with differential privacy. In: Proceedings of the 2016 ACM SIGSAC conference on computer and communications security, pp 308–318. https://doi.org/10.1145/2976749.2978318
23. Chuanxin Z, Yi S, Degang W (2020) Federated learning with gaussian differential privacy. In: Proceedings of the 2020 2nd international conference on robotics, intelligent control and artificial intelligence, pp 296–301. https://doi.org/10.1145/3438872.3439097
24. Ye Q, Meng X, Zhu M, Huo Z (2018) Survey on local differential privacy. J Softw 29(7):1981–2005. https://doi.org/10.13328/j.cnki.jos.005364
25. Papageorgiou A, Strigkos M, Politou E et al (2018) Security and privacy analysis of mobile health applications: the alarming state of practice. IEEE Access 6:9390–9403. https://doi.org/10.1109/ACCESS.2018.2799522
26. Duchi JC, Jordan MI, Wainwright MJ (2013) Local privacy and statistical minimax rates. In: 2013 IEEE 54th annual symposium on foundations of computer science. IEEE, pp 429–438. https://doi.org/10.1109/FOCS.2013.53
27. Wang T, Zhang X, Feng J et al (2020) A comprehensive survey on local differential privacy toward data statistics and analysis. Sensors 20(24):7030. https://doi.org/10.3390/s20247030
28. Greenberg A (2016) Apple's 'differential privacy' is about collecting your data---but not your data. https://www.wired.com/2016/06/apples-differential-privacy-collecting-data/. Accessed 13 June 2016
29. Erlingsson Ú, Pihur V, Korolova A (2014) Rappor: randomized aggregatable privacy-preserving ordinal response. In: Proceedings of the 2014 ACM SIGSAC conference on computer and communications security, pp 1054–1067. https://doi.org/10.1145/2660267.2660348
30. Osia SA, Shamsabadi AS, Taheri A et al (2017) Privacy-preserving deep inference for rich user data on the cloud. arXiv:1710.01727. https://doi.org/10.48550/arXiv.1710.01727
31. Tran L, Kong D, Jin H, Liu J (2016) Privacy-cnh: A framework to detect photo privacy with convolutional neural network using hierarchical features. In: Thirtieth AAAI conference on artificial intelligence, vol 30, no 1. https://doi.org/10.1609/aaai.v30i1.10169
32. Yu J, Zhang B, Kuang Z, Lin D, Fan J (2016) iPrivacy: image privacy protection by identifying sensitive objects via deep multi-task learning. IEEE Trans Inf Forensics Secur 12(5):1005–1016. https://doi.org/10.1109/TIFS.2016.2636090
33. Shokri R, Shmatikov V (2015) Privacy-preserving deep learning. In: Proceedings of the 22nd ACM SIGSAC conference on computer and communications security, pp 1310–1321. https://doi.org/10.1145/2810103.2813687
34. Yu L, Liu L, Pu C, Gursoy ME, Truex S (2019) Differentially private model publishing for deep learning. In: 2019 IEEE symposium on security and privacy (SP). IEEE, pp 332–349. https://doi.org/10.1109/SP.2019.00019
35. Nasr M, Shokri R (2020) Improving deep learning with differential privacy using gradient encoding and denoising. arXiv preprint arXiv:2007.11524. https://doi.org/10.48550/arXiv.2007.11524
36. Jayaraman B, Evans D (2019) Evaluating differentially private machine learning in practice. In: 28th USENIX security symposium (USENIX security 19), pp 1895–1912. https://doi.org/10.48550/arXiv.1902.08874
37. Blanco-Justicia A, Sánchez D, Domingo-Ferrer J et al (2022) A critical review on the use (and misuse) of differential privacy in machine learning. ACM Comput Surv 55(8):1–16. https://doi.org/10.1145/3547139
38. Papernot N, Thakurta A, Song S, Chien S, Erlingsson Ú (2020) Tempered sigmoid activations for deep learning with differential privacy. arXiv:2007.14191. https://doi.org/10.1609/aaai.v35i10.17123
39. Ayday E, Hubaux JP (2016) Privacy and security in the genomic era. In: Proceedings of the 2016 ACM SIGSAC conference on computer and communications security, pp 1863–1865. https://doi.org/10.1145/2976749.2976751
40. Raisaro JL, Ayday E, Hubaux JP (2014) Patient privacy in the genomic era. Praxis 103(10):579–586. https://doi.org/10.1024/1661-8157/a001657

41. Naveed M, Ayday E, Clayton EW et al (2015) Privacy in the genomic era. ACM Comput Surv 48(1):1–44. https://doi.org/10.1145/2767007
42. He Z, Li Y, Li J et al (2017) Addressing the threats of inference attacks on traits and genotypes from individual genomic data. In: 13th international symposium bioinformatics research and applications, pp 223–233. https://doi.org/10.1007/978-3-319-59575-7_20
43. Johnson A, Shmatikov V (2013) Privacy-preserving data exploration in genome-wide association studies. In: Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining, pp 1079–1087. https://doi.org/10.1145/2487575.2487687
44. Yu F, Fienberg SE, Slavković AB et al (2014) Scalable privacy-preserving data sharing methodology for genome-wide association studies. J Biomed Inform 50:133–141. https://doi.org/10.1016/j.jbi.2014.01.008
45. Humbert M, Ayday E, Hubaux JP et al (2014) Reconciling utility with privacy in genomics. In: Proceedings of the 13th workshop on privacy in the electronic society, pp 11–20. https://doi.org/10.1145/2665943.2665945
46. Tramèr F, Huang Z, Hubaux JP et al (2015) Differential privacy with bounded priors: reconciling utility and privacy in genome-wide association studies. In: Proceedings of the 22nd ACM SIGSAC conference on computer and communications security, pp 1286–1297. https://doi.org/10.1145/2810103.2813610
47. He Z, Li Y, Li J et al (2018) Achieving differential privacy of genomic data releasing via belief propagation. Tsinghua Sci Technol 23(4):389–395. https://doi.org/10.26599/TST.2018.9010037
48. Almadhoun N, Ayday E, Ulusoy Ö (2020) Inference attacks against differentially private query results from genomic datasets including dependent tuples. Bioinformatics 36:i136–i145. https://doi.org/10.1093/bioinformatics/btaa475
49. Almadhoun N, Ayday E, Ulusoy Ö (2020) Differential privacy under dependent tuples—the case of genomic privacy. Bioinformatics 36(6):1696–1703. https://doi.org/10.1093/bioinformatics/btz837
50. Yilmaz E, Ji T, Ayday E et al (2022) Genomic data sharing under dependent local differential privacy. In: Proceedings of the twelfth ACM conference on data and application security and privacy, pp 77–88. https://doi.org/10.1145/3508398.3511519
51. Chen J, Wang WH, Shi X (2020) Differential privacy protection against membership inference attack on machine learning for genomic data. Biocomputing 2021:26–37. https://doi.org/10.1142/9789811232701_0003
52. Hu Y, Ge L, Zhang G, Qin D (2019) Research on differential privacy for medical health big data processing. In: 2019 20th international conference on parallel and distributed computing, applications and technologies (PDCAT). IEEE, pp 140–145. https://doi.org/10.1109/PDCAT46702.2019.00036
53. Tu Z, Liu S, Xiong X, Zhao J, Cai Z (2020) Differential private average publishing of numerical stream data for wearable devices. J Comput Appl 40(6):6. https://doi.org/10.11772/j.issn.1001-9081.2019111929
54. Kim JW, Jang B, Yoo H (2018) Privacy-preserving aggregation of personal health data streams. PLoS ONE 13(11):e0207639. https://doi.org/10.1371/journal.pone.0207639
55. Li Z, Wang B, Li J, Hua Y, Zhang S (2022) Local differential privacy protection for wearable device data. PLoS ONE 17(8):e0272766. https://doi.org/10.1371/journal.pone.0272766
56. Han S, Zhao S, Li Q et al (2015) PPM-HDA: privacy-preserving and multifunctional health data aggregation with fault tolerance. IEEE Trans Inf Forensics Secur 11(9):1940–1955. https://doi.org/10.1109/TIFS.2015.2472369
57. Lin C, Wang P, Song H et al (2016) A differential privacy protection scheme for sensitive big data in body sensor networks. Ann Telecommun 71:465–475. https://doi.org/10.1007/s12243-016-0498-7
58. Hadian M, Liang X, Altuwaiyan T et al (2016) Privacy-preserving mhealth data release with pattern consistency. In: 2016 IEEE global communications conference (GLOBECOM). IEEE, pp 1–6. https://doi.org/10.1109/GLOCOM.2016.7842173
59. Bozkir E, Günlü O, Fuhl W et al (2021) Differential privacy for eye tracking with temporal correlations. PLoS ONE 16(8):e0255979. https://doi.org/10.1371/journal.pone.0255979
60. Wu G, Wang S, Ning Z et al (2021) Privacy-preserved electronic medical record exchanging and sharing: a blockchain-based smart healthcare system. IEEE J Biomed Health Inform 26(5):1917–1927. https://doi.org/10.1109/JBHI.2021.3123643
61. Chen S, Fu A, Yu S et al (2021) DP-QIC: a differential privacy scheme based on quasi-identifier classification for big data publication. Soft Comput 25:7325–7339. https://doi.org/10.1007/s00500-021-05692-7

62. Zhang S, Li X (2022) Differential privacy medical data publishing method based on attribute correlation. Sci Rep 12(1):15725. https://doi.org/10.1038/s41598-022-19544-3
63. Ziller A, Usynin D, Braren R et al (2021) Medical imaging deep learning with differential privacy. Sci Rep 11(1):1–8. https://doi.org/10.1038/s41598-021-93030-0
64. Yuan D, Zhu X, Wei M et al (2019) Collaborative deep learning for medical image analysis with differential privacy. In: 2019 IEEE global communications conference (GLOBECOM). IEEE, pp 1–6. https://doi.org/10.1109/GLOBECOM38437.2019.9014259
65. Adnan M, Kalra S, Cresswell JC et al (2022) Federated learning and differential privacy for medical image analysis. Sci Rep 12(1):1953. https://doi.org/10.1038/s41598-022-05539-7
66. Gao Y, Zhang P, Zhou C et al (2023) HGNAS++: efficient architecture search for heterogeneous graph neural networks. IEEE Trans Knowl Data Eng. https://doi.org/10.1109/TKDE.2023.3239842
67. Gao Y, Zhang P, Yang H et al (2022) GraphNAS++: distributed architecture search for graph neural networks. IEEE Trans Knowl Data Eng. https://doi.org/10.1109/TKDE.2022.3178153
68. Zheng Z, Wang C, Xu T et al (2021) Drug package recommendation via interaction-aware graph induction. In: Proceedings of the web conference 2021, pp 1284–1295. https://doi.org/10.1145/3442381.3449962
69. Shen ZA, Luo T, Zhou YK et al (2021) NPI-GNN: predicting ncRNA–protein interactions with deep graph neural networks. Brief Bioinform 22(5):bbab051. https://doi.org/10.1093/bib/bbab051
70. Réau M, Renaud N, Xue LC et al (2023) DeepRank-GNN: a graph neural network framework to learn patterns in protein–protein interfaces. Bioinformatics 39(1):btac759. https://doi.org/10.1093/bioinformatics/btac759
71. Wei Y, Fu X, Sun Q et al (2022) Heterogeneous graph neural network for privacy-preserving recommendation. arXiv:2210.00538. https://doi.org/10.48550/arXiv.2210.00538
72. Sajadmanesh S, Gatic-Perez D (2021) Locally private graph neural networks. In: Proceedings of the 2021 ACM SIGSAC conference on computer and communications security, pp 2130–2145. https://doi.org/10.1145/3460120.3484565
73. Ge YF, Orlowska M, Cao J et al (2022) MDDE: multitasking distributed differential evolution for privacy-preserving database fragmentation. VLDB J 31(5):957–975. https://doi.org/10.1007/s00778-021-00718-w
74. Ge YF, Orlowska M, Cao J et al (2021) Knowledge transfer-based distributed differential evolution for dynamic database fragmentation. Knowl Based Syst 229:107325. https://doi.org/10.1016/j.knosys.2021.107325
75. Ge YF, Yu WJ, Cao J et al (2020) Distributed memetic algorithm for outsourced database fragmentation. IEEE Trans Cybern 51(10):4808–4821. https://doi.org/10.1109/TCYB.2020.3027962

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.