Original Paper

Evaluation of Privacy Risks of Patients' Data in China: Case Study

Mengchun Gong[1*], MD; Shuang Wang[2*], PhD; Lezi Wang[1], MSc; Chao Liu[1], PhD; Jianyang Wang[3], MD; Qiang Guo[4], MSc; Hao Zheng[5], PhD; Kang Xie[6], PhD; Chenghong Wang[3], MSc; Zhouguang Hui[3,7], MD

[1]Digital China Health Technologies Corporation Limited, Beijing, China

[2]Shanghai Putuo People's Hospital, Tongji University, Shanghai, China

[3]Department of Radiation Oncology, National Cancer Center/National Clinical Research Center for Cancer/Cancer Hospital, Chinese Academy of Medical Sciences and Peking Union Medical College, Beijing, China

[4]Big Data Center, National Cancer Center/National Clinical Research Center for Cancer/Cancer Hospital, Chinese Academy of Medical Sciences and Peking Union Medical College, Beijing, China

[5]Department of Bioinformatics, Hangzhou Nuowei Information Technology, Hangzhou, China

[6]The Third Research Institute of Ministry of Public Security, Key Lab of Information Network Security, Ministry of Public Security, Shanghai, China

[7]Department of VIP Medical Services, National Cancer Center/National Clinical Research Center for Cancer/Cancer Hospital, Chinese Academy of Medical Sciences and Peking Union Medical College, Beijing, China

[*]these authors contributed equally

Corresponding Author:
Zhouguang Hui, MD
Department of VIP Medical Services
National Cancer Center/National Clinical Research Center for Cancer/Cancer Hospital, Chinese Academy of Medical Sciences and Peking Union Medical College
Panjiayuan Nanli #17, Chaoyang District
Beijing, 100021
China
Phone: 86 010 87787656
Email: huizg@cicams.ac.cn

Abstract

Background: Patient privacy is a ubiquitous problem around the world. Many existing studies have demonstrated the potential privacy risks associated with sharing of biomedical data. Owing to the increasing need for data sharing and analysis, health care data privacy is drawing more attention. However, to better protect biomedical data privacy, it is essential to assess the privacy risk in the first place.

Objective: In China, there is no clear regulation for health systems to deidentify data. It is also not known whether a mechanism such as the Health Insurance Portability and Accountability Act (HIPAA) safe harbor policy will achieve sufficient protection. This study aimed to conduct a pilot study using patient data from Chinese hospitals to understand and quantify the privacy risks of Chinese patients.

Methods: We used g-distinct analysis to evaluate the reidentification risks with regard to the HIPAA safe harbor approach when applied to Chinese patients' data. More specifically, we estimated the risks based on the HIPAA safe harbor and limited dataset policies by assuming an attacker has background knowledge of the patient from the public domain.

Results: The experiments were conducted on 0.83 million patients (with data field of *date of birth, gender, and surrogate ZIP codes* generated based on home address) across 33 provincial-level administrative divisions in China. Under the Limited Dataset policy, 19.58% (163,262/833,235) of the population could be uniquely identifiable under the g-distinct metric (ie, 1-distinct). In contrast, the Safe Harbor policy is able to significantly reduce privacy risk, where only 0.072% (601/833,235) of individuals are uniquely identifiable, and the majority of the population is 3000 indistinguishable (ie the population is expected to share common attributes with 3000 or less people).

Conclusions: Through the experiments based on real-world patient data, this work illustrates that the results of g-distinct analysis about Chinese patient privacy risk are similar to those from a previous US study, in which data from different organizations/regions might be vulnerable to different reidentification risks under different policies. This work provides reference to Chinese health care entities for estimating patients' privacy risk during data sharing, which laid the foundation of privacy risk study about Chinese patients' data in the future.

(JMIR Med Inform 2020;8(2):e13046) doi: 10.2196/13046

KEYWORDS

patient privacy; privacy risk; Chinese patients' data; data sharing; re-identification

Introduction

Background

Medical data are naturally distributed across institutions as patients might visit different hospitals at different times or for different diseases. To better understand the risk factors and efficacy of treatment, it is necessary to share data and analyze them. However, patient data are highly sensitive as they contain medical and personal identity information [1-5]. This is a ubiquitous problem. China has the largest population in the world, and the issue of privacy is becoming a big concern for the health care system to share medical data. Inappropriate handling of these sensitive data can lead to privacy leakage, which in turn can result in social embarrassment and commercial fraudulence [6-10].

In the United States, the Health Insurance Portability and Accountability Act (HIPAA) [11] safeguards the health care data. Thus, protected health information can only be considered as deidentified if it is sanitized by one of the following approaches specified by the HIPAA privacy rule [12]: (1) expert determination or (2) safe harbor. The first approach is to recruit an expert with appropriate knowledge and experience to render information with minimal risk to be reidentified. The second approach is to use the safe harbor approach, which explicitly denotes 18 identifiers that need to be removed. The average fine levied for a HIPAA breach is between US $10,000 and US $50,000 per medical record. There are similar guidelines in other countries, for example, the European Union's General Data Protection Regulation [13] and Canada's Personal Information Protection and Electronic Documents Act [14], which regulate patient records and other sensitive information. In South Korea (Korea) and Japan, the general law regulating privacy and data protection is the Personal Information Protection Act, and there is a more complete list of international privacy-related laws by country and region.

In China, the *Network Security Law of the People's Republic of China* [15], which was formally put into effect on June 1, 2017, regulates that network providers must not disclose, falsify, or destroy any personal information they have collected. Any network provider must not disclose this personal information to any third party without obtaining consent from data owners, except for the data that cannot be used to reidentify a specific individual. However, there are no guidelines on how personal information can be processed to satisfy the above regulation. On December 29, 2017, the Chinese government formally released a new regulation called *Information Security Technology and Personal Information Security Specification* (referred to as *Specification*) [16]. In the *Specification*, the Chinese government has clearly defined privacy-related terms such as "personal information controller," "collection," "informed consent," "user portrait," "personal information security impact assessment," "deletion," and "deidentification." The *Specification* also defines security requirements for different phases (eg, collection, storage, processing, transfer, and disclosure) in handling personal data. However, the *Specification* also has several limitations. First, the *Specification* is a recommended national standard and not a legal regulation; thus, it might not be stringently enforced by different entities. Second, the *Specification* mainly focuses on general purpose information security, where no specific guidance is provided for tackling medical or health care data. For example, in the *Specification*, almost all medical-related data are defined as highly sensitive data. The *Specification*, on one hand, emphasizes the importance of obtaining explicit consent of individuals when collecting, using, or disclosing sensitive personal information, whereas, on the other hand, there are several situations have been added as exceptions. For instance, if the personal information controller is an academic research institution, then it is necessary for them to perform statistical or academic research in public interest. If they provide external academic research or description results with deidentified personal information, they will be exempted from obtaining explicit consent from each individual. In addition, if the use of personal data is directly related to public safety, public health, and major public interest, then there is no need to obtain individual consent on personal data usage. In the third case, if there are certain difficulties in obtaining personal consent and if the use of personal data is to safeguard the major legal rights such as the life and property of the subject or individuals, then such usage of personal information will be exempted from obtaining explicit consent. In the *Specification*, deidentification is defined as a process by which the personal information is technically marked out so that the remaining information cannot be used to reidentify the individual without using additional information. On August 15, 2017, *Information Security Technology and Personal Information Deidentification Specification* was published by the Chinese government for public comments, which also introduced many existing deidentification procedures. However, there is still no clear guidance in China about how to deidentify health care data to ensure sufficient protection of the privacy of individual patients. Owing to the difference in population density, it is also not clear if similar protection mechanisms such as the HIPAA safe harbor rule will provide comparable protection to the Chinese patients' data. There is also a difference between the external sources of background information that attackers can leverage. For example, there is no public voter's registry in China, but social networks make public a considerable amount of demographic information of their users such as gender, birthday, school, and job. It is necessary to measure the privacy risks of Chinese patients' data to better understand the associated privacy risk.

Besides direct identifiers (such as name, national ID, and address), the privacy risk of a medical record is related to the rareness of its key variable values. For example, if there is a unique combination of birthday, gender, and ZIP code, the corresponding record is more likely to be reidentified when compared with records that have duplicated characteristics in the database. It has been shown in the study by Sweeney [17]

that 87% of the US population can be uniquely identified by the triplet (birthday, gender, and ZIP code), which reveals a high privacy risk if data are shared without sanitization. It is important to measure the rareness of individual records in a database to understand the potential risk it carries.

Privacy risk measurements and anonymization methods such as *k*-anonymity [18], *l*-diversity [19], *t*-closeness [20], and differential privacy (DP) [21] have motivated many algorithmic and theoretical studies. *k*-anonymity reduces the granularity of data representation using data generalization and suppression technologies. The parameter *k* indicates the number of records within the equivalence class, in which an adversary cannot distinguish an individual. A larger *k* implies a smaller reidentification risk. El Emam et al [22] applied an optimized *k*-anonymity algorithm for health data deidentification. *l*-diversity improves *k*-anonymity by ensuring that the intragroup diversity for sensitive values is controlled by the parameter *l* [19]. *t*-closeness provides a stronger privacy notion than *l*-diversity, where *t*-closeness requires that the distribution of a private attribute in any equivalence class is computationally indistinguishable (ie, no larger than *t*) from the distribution of the attribute in the overall table. Both techniques have been adopted in many medical data deidentification applications [23]. Recently, DP became one of the de facto standards for achieving strong privacy guarantees, which assumes that an attacker with any background knowledge cannot tell if a particular individual's information has been included or not based on the differentially private outputs [24]. DP technology has also been applied to protect health care data dissemination and analysis [25-27]. In this work, we were interested in a measurement to evaluate the reidentification risks with respect to the HIPAA privacy rule when applied to Chinese patients' data. However, none of the aforementioned methods can be directly adopted for serving this goal. Therefore, this work resorts to the g-distinct method previously proposed in the study by Malin et al [28] for evaluating reidentification risks of HIPAA-deidentified data in the United States.

Objectives

The main objectives and contributions of this work are three-fold: (1) to provide one of the first large-scale studies on the privacy risks of Chinese patients' data, (2) to design specifically experimental studies based on the characteristics of Chinese patients' data for evaluating the patient privacy risks in China, and (3) to provide references for improving the current privacy protection and rulemaking for Chinese patients' data.

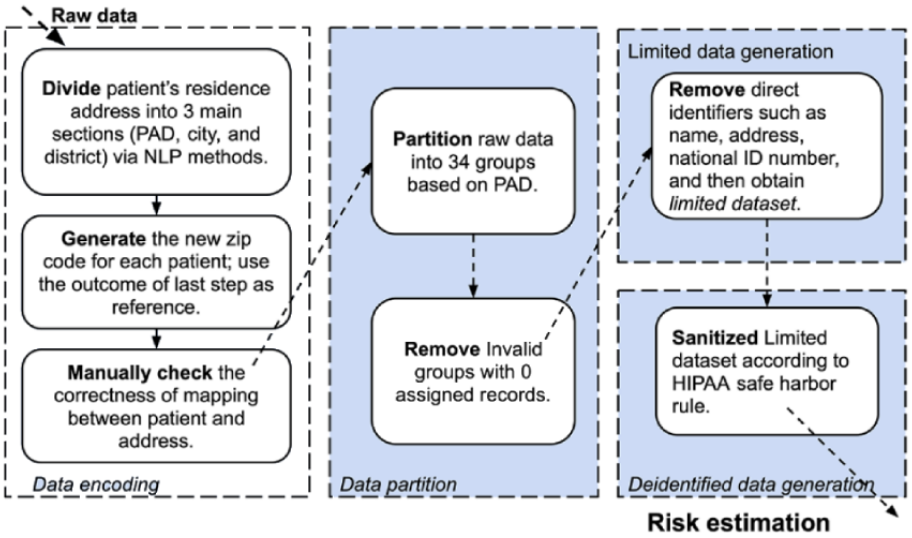
Methods

Data Preprocessing

The datasets used for conducting our experiments are based on cancer patients' records in the Malignant Cancer Big-data Processing Analysis and Application Research Project (MCBPAARP), which is supported by the National Cancer Institute's High-Tech Research and Development *863* program. The *863* program was led by the Ministry of Science and Technology of the People's Republic of China, where the goal of this program is to promote the development of advanced technologies across different fields. This study under MCBPAARP has been approved by the Institutional Review Board of National Cancer Center/National Cancer Hospital and Chinese Academy of Medical Sciences, also known as Ethics Committee of National Cancer Center, National Cancer Hospital, and Chinese Academy of Medical Sciences of Peking Union Medical College under the project ID 2017YFC1311000. In China, hospitals have to update inpatients' medical record home page to National Health Commission of the People's Republic of China under a unified standard. The Chinese patients' data attributes used in this study include fields P5 gender, P6 birthday, and P801 home address (used to generate masked ZIP codes).

Figure 1 illustrates the methods used for the raw data preprocessing in this study for privacy risk analysis, which includes four phases: (1) data encoding; (2) data partition; (3) limited dataset generation; and (4) *deidentified* dataset generation. The phases are described as follows:

Figure 1. The workflow of raw data preprocessing in this study. HIPAA: Health Insurance Portability and Accountability Act; NLP: neurolinguistic programming; PAD: provincial-level administrative division.



Data Encoding

In Chinese patients' data, the quality of ZIP codes from patients' raw data is extremely low, which may be either missing or too generalized (ie, only at city level). To overcome this problem, we introduced the following encoding scheme to convert the patient's address information into geocodes as surrogate ZIP codes in this study. We first divided the patient's residence address into three sections (ie, provincial-level administrative divisions [PADs], city, and district) by using natural language processing methods. Thereafter, we encoded PAD, city, and district with 2 digits, 3 digits, and 4 digits, respectively, which resulted in surrogate ZIP codes for a total of 9 digits. To ensure the data quality, we conducted two rounds of manual checking for the mapping correctness between the patient's residence addresses and their surrogate ZIP codes. We excluded patients with missing residence address information and the records with obvious logical error (ie, the patient's date of birth [DOB] is 1900-01-01). Finally, 0.83 million hospitalized patients' medical records were selected in this study.

Data Partition

We partitioned raw patient data into different groups based on their PADs. Through this phase, we ended up with 33 nonempty PADs except for Hong Kong PAD (see Multimedia Appendix 1 for more details of PADs).

Limited Dataset Generation

After the data encoding phase, we further removed additional explicit identifiers, such as name, address, and national ID number from the raw data, which left us with the *limited dataset* with only DOB, gender, and surrogate ZIP codes.

Deidentified Dataset Generation

On the basis of the HIPAA safe harbor rule, we further sanitized the limited dataset by generalizing DOB to year and all surrogate ZIP codes to the first 6 digits.

Risk Evaluation

To evaluate the privacy risk of the preprocessed Chinese patients' data, we adopted the g-distinct method introduced in the study by Malin et al [28] for studying a similar problem in the United States. The g-distinct method quantifies the uniqueness of individual records within a database, where an individual is said to be unique if such an individual has a combination of personal attributes that no other individuals in the same dataset has. Furthermore, we say an individual is g-distinct if the combination of their attributes is identical to at the most g-1 other individuals in the whole dataset space. For example, suppose an individual has the following combination of attributes: age at 35 years and gender as male. If there does not exist any other individual whose age and gender are also 35 years and male, respectively, then such an individual is considered as unique (ie, 1-distinct). In addition, if the total number of individuals with the same combination of attributes is equal to *k*, then we state this individual is *k-distinct*.

In other words, we can also describe the g-distinct as the sum over the number of patients in all bins with less than or equal to *g* individuals, which can be written as shown in equation (1):

$$h(g)=\sum_{i=1}^{g}i|bin(i)| \quad (1)$$

Here *g* denotes the parameter and |bin(i)| refers to the number of bins with exact *i* patients having identical attributes. This measurement serves as a proxy to the risk of stratified population with different combinations of characteristics. In this study, we applied the above *g*-distinct metric to the Chinese patients' data to evaluate the privacy risk.

Results

Experimental Setup

The g-distinct analysis is a population inspection method that allows us to investigate a particular cross-section for specific population collection. Such particular cross-section represents the set of individuals whose private records are most vulnerable to reidentification attacks. In our experiments, we conducted g-distinct analysis over the limited dataset (ie, DOB, gender, and ZIP code) and the safe harbor dataset (ie, birth year, gender, and masked ZIP code) to examine how the safe harbor data can improve the privacy of individual patients over limited data.

Experimental Results

The results of g-distinct analysis based on nationwide datasets for both the safe harbor dataset and limited dataset are illustrated in Figure 2. In Figure 2, the left and right subgraphs represent the g-distinct analysis results for limited and safe harbor datasets, respectively. According to the nationwide g-distinct analysis results, we have two major observations. On the one hand, without sufficient deidentification process (ie, the limited data on the left), the whole dataset is highly risky. For instance, 19.58% (163,262/833,235) of the population is 1-distinct in the limited dataset (ie, uniquely identifiable under the g-distinct metric). In addition, more than 90.6% of the population is 10-distinct, which implies that the majority of the population in the limited dataset is expected to share common attributes with 10 or less people. Such sheer number of distinct individuals results in a huge difficulty for privacy protection. Thus, in such cases, the limited data are extremely vulnerable to reidentification attacks. On the other hand, as shown in Figure 2, the safe harbor dataset is able to significantly preserve the patient's privacy, in which only 0.072% (601/833,235) of individuals are uniquely identifiable (ie, 1-distinct), and the majority of the population (around 95%) is 3000 indistinguishable.

We also studied the relationship between distinct individuals and the underlying populations, which simulates the impact of different ZIP code–masking strategies on the privacy protection. The results of this experiment have been illustrated in Figure 3. There are a total of 34 PADs in China. As there were no patient records with residence address within Hong Kong in the collected datasets, we estimated the percentage of 1-distinct population over the other 33 PADs (see Multimedia Appendix 2 for more details). Figure 3 shows the percentage of 1-distinct population associated with each PAD in an ascending order of the sample population in the given PAD. The 2 subgraphs are the results over limited dataset and the safe harbor dataset, respectively. Owing to the accommodation of different scales of 1-distinct percentage along with the increase in population,

the 2 plots are depicted in log-log scale. As shown in Figure 3, both results show a similar tendency, where the percentage of 1-distinct population decreases as the sampled population increases in different PADs. This is because a PAD with more sampled population tends to result in a higher probability to have more than 1 patient who shares the same attributes. Another observation from the result is that when the sampled population increases, the percentage of 1-distinct population of the safe harbor dataset decreases dramatically. When the population has increased to 10,000, the 1-distinct percentage has already dropped to 0.05%. In contrast, the decreasing tendency for the limited dataset seems more moderate (ie, potentially higher privacy risks). We can see that there is still 5% population more with 1-distinct for a PAD of 200,000 population in the limited dataset.

Figure 2. The g-distinct versus percentage of population under limited dataset and safe harbor dataset, respectively.

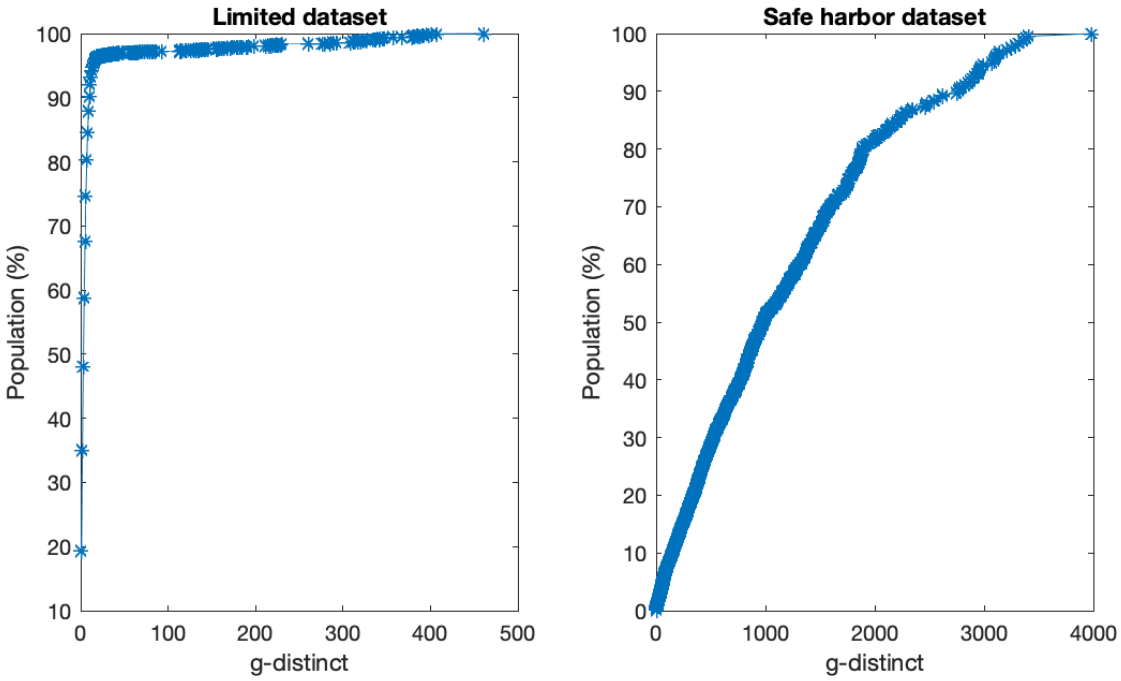
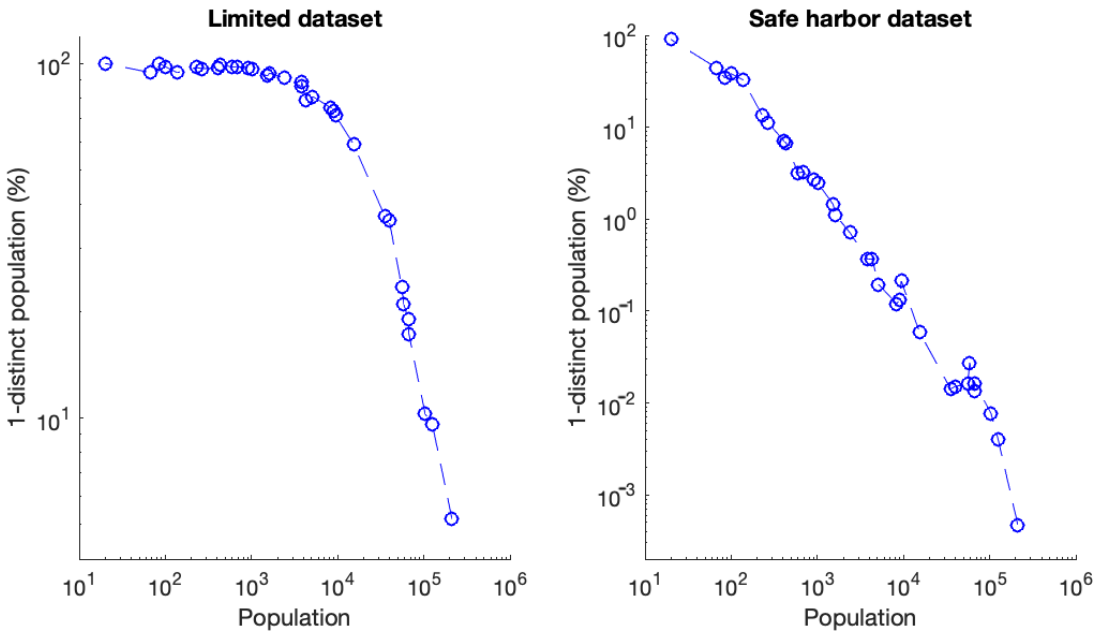


Figure 3. Percentage of 1-distinct versus total population under limited dataset and safe harbor dataset, respectively.



Discussion

Principal Findings

This work is one of the first large-scale studies on evaluating privacy risks of Chinese patients' data, which analyzed the reidentification risks based on the HIPAA safe harbor and limited dataset policies. The originality of this work can be summarized in three ways:

1. Originality in exploring new observations: Although many Chinese Acts [15] and national specification/regulations [16] have cited the HIPAA safe harbor rule [11] as a reference standard for guiding patient data deidentification in China, there is still a lack of quantitative observation on the reidentification risk of Chinese patient data when applying the HIPAA safe harbor standard. This work provides one of the first quantitative studies on large-scale nationwide Chinese patients' data toward this goal.
2. Originality in designing new experiments: Some Chinese patients' data attributes are unique and different from those in the United States. Therefore, these data cannot be applied directly to the risk assessment method used in previous US studies. For example, Chinese patients' data typically have extremely low quality of ZIP codes, which may be either missing or too generalized (ie, only at city level). Thus, we designed new data encoding, data partition, and data masking schemes based on Chinese patients' data characteristics to meet this goal.
3. Originality in contributing new knowledge: We made an assumption that the risk evaluation scheme defined by the HIPAA is satisfactory with respect to Chinese patient data as well. According to this assumption, we designed and implemented our experimental studies based on Chinese patients' data. As patient privacy protection is a very important topic, many other research studies have been conducted in Europe [29-31], Japan [32,33], and Australia [34]. However, most of these studies are more qualitative in orientation and usually not suitable for Chinese patients' data, which mainly focus on the interpretation and comparison of laws and regulations. In contrast, the focus of this work was to quantify the Chinese patient privacy risks with large-scale and real-world patient data collected from China. According to our experimental studies, our assumption is supported by the results of this work, which illustrates findings similar to those of a previous US study by Malin et al [28] that evaluated reidentification of US patient data associated with the HIPAA policies. Such studies are amenable to various kinds of meta-evaluations, enabling administrative roles such as government's policy makers and datacenter administrators to be able to evaluate policies and to determine the potential impact on reidentification risk. The experimental results demonstrate the power of the g-distinct analysis applied on Chinese patients' data. In general, according to the experimental results, the safe harbor dataset provides much stronger privacy protection in terms of 1-distinct than that provided by the limited dataset in Chinese patients' data.

Limitations

In general, this work provides justification for reidentification risk estimates on Chinese patient records before sharing data. However, the proposed studies still have a few limitations. First, the privacy risk that we estimated for the case study is based on the cancer patients' data without including patients with other diseases. Second, although the datasets are from 33 of 34 PADs, the study is still limited by the data scale, which only covers less than 0.06% of the Chinese population (ie, 0.83 million patients' records vs 1.5 billion total population). Third, the demographic information used in this study is also limited. For example, it is unfeasible to measure the identifiability based on nationality. Therefore, raw data are collected with selection bias because of aforementioned limitations. All these limitations justify further investigation along this line.

Conclusions

The study of Chinese patients' privacy risk in this work fills the gap of the privacy research between the United States and China. Moreover, as the Chinese government has not yet issued specific regulations or policies directly against privacy protection of citizens' health data, our experimental studies have the potential for Chinese officials to improve current health data–sharing regulations. The policy might vary largely among provinces, as according to the statistics, the g-distinct measurements vary widely across the provinces as well. Privacy officials might issue flexible policies for different regions.

Acknowledgments

This study is supported by a national key research and development program (2017YFC1311000 2017YFC1311001). This work is partially supported by Key Lab of Information Network Security of Ministry of Public Security (The Third Research Institute of Ministry of Public Security) under funding C19609. The authors would like to thank Dr Hua Xu and Dr Xiaoqian Jiang for the helpful discussions and suggestions.

Authors' Contributions

MG and SW share first authorship and contributed the majority of the writing and conducted major parts of the methods and experiment design, where a part of this work was done by SW in the Institutes for Systems Genetics West China Hospital. LW and CL conducted some experiments and contributed to data preprocessing and paper writing. HZ and KX contributed to paper writing and provided comments on methods. JW, QG, and ZH provided the motivation for this work, data collection, paper writing, detailed edits, and critical suggestions. All authors reviewed the final manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1

Abbreviations of provinces in China.
[DOCX File , 15 KB-Multimedia Appendix 1]

Multimedia Appendix 2

The g-distinct analysis results.
[DOCX File , 18 KB-Multimedia Appendix 2]

References

1. Jiang Y, Hamer J, Wang C, Jiang X, Kim M, Song Y, et al. SecureLR: Secure Logistic Regression Model via a Hybrid Cryptographic Protocol. IEEE/ACM Trans. Comput. Biol. and Bioinf 2019 Jan 1;16(1):113-123. [doi: 10.1109/tcbb.2018.2833463]
2. Chen F, Wang C, Dai W, Jiang X, Mohammed N, Al Aziz MM, et al. PRESAGE: PRivacy-preserving gEnetic testing via SoftwAre Guard Extension. BMC Med Genomics 2017 Jul 26;10(S2). [doi: 10.1186/s12920-017-0281-2]
3. Wang X, Tang H, Wang S, Jiang X, Wang W, Bu D, et al. iDASH secure genome analysis competition 2017. BMC Med Genomics 2018 Oct 11;11(S4). [doi: 10.1186/s12920-018-0396-0]
4. Chenghong W, Jiang Y, Mohammed N, Chen F, Jiang X, Al Aziz MM, et al. SCOTCH: Secure Counting Of encrypTed genomiC data using a Hybrid approach. AMIA Annu Symp Proc 2017;2017:1744-1753 [FREE Full text] [Medline: 29854245]
5. Chen F, Wang S, Jiang X, Ding S, Lu Y, Kim J, et al. PRINCESS: Privacy-protecting Rare disease International Network Collaboration via Encryption through Software guard extensionS. Bioinformatics 2017 Mar 15;33(6):871-878 [FREE Full text] [doi: 10.1093/bioinformatics/btw758] [Medline: 28065902]
6. Krishnamurthy B, Wills CE. Privacy Leakage in Mobile Online Social Networks. 2010 Jun 10 Presented at: The 3rd Wonference on Online Social Networks; 2010; Berkeley, California URL: https://www.usenix.org/legacy/events/wosn10/tech/full_papers/Krishnamurthy.pdf
7. Ignatenko T, Willems FMJ. Privacy leakage in biometric secrecy systems. 2008 Jun 10 Presented at: 46th Annual Allerton Conference on Communication, Control, and Computing; 2008; Allerton. [doi: 10.1109/ALLERTON.2008.4797752]
8. Gymrek M, McGuire AL, Golan D, Halperin E, Erlich Y. Identifying personal genomes by surname inference. Science 2013 Jan 18;339(6117):321-324 [FREE Full text] [doi: 10.1126/science.1229566] [Medline: 23329047]
9. Hansson MG, Lochmüller H, Riess O, Schaefer F, Orth M, Rubinstein Y, et al. The risk of re-identification versus the need to identify individuals in rare disease research. Eur J Hum Genet 2016 Nov 25;24(11):1553-1558 [FREE Full text] [doi: 10.1038/ejhg.2016.52] [Medline: 27222291]
10. Vaidya J, Shafiq B, Jiang X, Ohno-Machado L. Identifying inference attacks against healthcare data repositories. AMIA Jt Summits Transl Sci Proc 2013;2013:262-266 [FREE Full text] [Medline: 24303279]
11. Public Law. 1996. Health insurance portability and accountability act of 1996 URL: https://www.govinfo.gov/content/pkg/PLAW-104publ191/pdf/PLAW-104publ191.pdf [accessed 2020-01-17]
12. Summary of the HIPAA privacy rule. 2003. Others URL: https://www.hhs.gov/sites/default/files/privacysummary.pdf [accessed 2020-01-17]
13. Voigt P, von dem Bussche A. The EU General Data Protection Regulation (GDPR): A Practical Guide. Cham, Switzerland: Springer; Jun 10, 2017.
14. Charnetski W, Flaherty P, Robinson J. The Personal Information Protection and Electronic Documents Act: A Comprehensive Guide. In: Canada Law Book. Toronto: Canada Law Book; Jun 10, 2001:2001.
15. Peng PL. Ordinance of the People's Republic of China on the Protection of Computer Information System Security. Chinese Law & Government 2014 Dec 07;43(5):12-16. [doi: 10.2753/clg0009-4609430501]
16. China Law Blog. 2018. China’s Personal Information Security Specification: Get Ready for May 1 URL: https://www.chinalawblog.com/2018/02/chinas-personal-information-security-specification-get-ready-for-may-1.html [accessed 2018-06-26]
17. Sweeney L. CiNii. 2000. Uniqueness of simple demographics in the US Population, in LIDAP-WP4 URL: https://ci.nii.ac.jp/naid/10020493621/ [accessed 2020-01-17]
18. Sweeney L. Achieving k-Anonymity Privacy Protection Using Generalization and Suppression. Int J Unc Fuzz Knowl Based Syst 2012 May 02;10(05):571-588 [FREE Full text] [doi: 10.1142/S021848850200165X]
19. Machanavajjhala A, Gehrke J, Kifer D. L-diversity: Privacy beyond k-anonymity. 2006 Presented at: 22nd International Conference on Data Engineering (ICDE'06); April 3-7, 2006; Atlanta, GA. [doi: 10.1109/icde.2006.1]

20. Li N, Li T, Venkatasubramanian S. t-closeness: Privacy beyond k-anonymity and l-diversity. 2007 Presented at: 2007 IEEE 23rd International Conference on Data Engineering; June 4, 2007; Istanbul, Turkey URL: http://ieeexplore.ieee.org/abstract/document/4221659/ [doi: 10.1109/icde.2007.367856]
21. Dwork C. Differential Privacy: A Survey of Results. In: Theory and Applications of Models of Computation. Berlin, Heidelberg, Germany: Springer; Jun 10, 2008:19.
22. El Emam K, Dankar F, Issa R. A globally optimal k-anonymity method for the de-identification of health data. J Am Med Inform Assoc 2009;16:82. [doi: 10.1197/jamia.m3144]
23. Gkoulalas-Divanis A, Loukides G, Sun J. Publishing data from electronic health records while preserving privacy: a survey of algorithms. J Biomed Inform 2014 Aug;50:4-19 [FREE Full text] [doi: 10.1016/j.jbi.2014.06.002] [Medline: 24936746]
24. Jiang X, Sarwate AD, Ohno-Machado L. Privacy Technology to Support Data Sharing for Comparative Effectiveness Research. Medical Care 2013;51:S58-S65. [doi: 10.1097/mlr.0b013e31829b1d10]
25. Jiang Y, Wang C, Wu Z, Du X, Wang S. Privacy-preserving biomedical data dissemination via a hybrid approach. AMIA Annu Symp Proc 2018;2018:1176-1185 [FREE Full text] [Medline: 30815160]
26. 26 DF, El EK. The application of differential privacy to health data. 2012 Presented at: Proceedings of the Joint EDBT/ICDT Workshops; March 2012; Berlin, Germany URL: https://dl.acm.org/citation.cfm?id=2320816 [doi: 10.1145/2320765.2320816]
27. Dankar F, El EK. Practicing differential privacy in health care: A review. Trans Data Priv 2013;6:67.
28. Benitez K, Malin B. Evaluating re-identification risks with respect to the HIPAA privacy rule. J Am Med Inform Assoc 2010;17(2):169-177 [FREE Full text] [doi: 10.1136/jamia.2009.000026] [Medline: 20190059]
29. Tucker K, Branson J, Dilleen M, Hollis S, Loughlin P, Nixon MJ, et al. Protecting patient privacy when sharing patient-level data from clinical trials. BMC Med Res Methodol 2016 Jul 08;16 Suppl 1(S1):77 [FREE Full text] [doi: 10.1186/s12874-016-0169-4] [Medline: 27410040]
30. Wierda E, Eindhoven D, Schalij M, Borleffs CJW, Amoroso G, van Veghel D, et al. Privacy of patient data in quality-of-care registries in cardiology and cardiothoracic surgery: the impact of the new general data protection regulation EU-law. Eur Heart J Qual Care Clin Outcomes 2018 Oct 01;4(4):239-245. [doi: 10.1093/ehjqcco/qcy034] [Medline: 30060178]
31. Deguara I. Thesynapse.net. 2018. Protecting patients' medical records under the GDPR URL: https://www.um.edu.mt/library/oar/bitstream/123456789/40287/1/The_Synapse%2c_17%282%29_-_A1.pdf [accessed 2020-01-17]
32. Kim J. J Marshall J Info Tech & Privacy L. 2015. Japanese and American Privacy Laws, Comparative Analysis URL: https://repository.jmls.edu/cgi/viewcontent.cgi?article=1782&context=jitpl [accessed 2020-01-17]
33. Yamamoto H. Use of personal information in medical research in Japan. The Lancet 2016 Oct;388(10055):1981-1982. [doi: 10.1016/s0140-6736(16)31867-0]
34. Williamson OD, Cameron PA, McNeil JJ. Medical registry governance and patient privacy. Medical Journal of Australia 2004 Aug 02;181(3):125-126. [doi: 10.5694/j.1326-5377.2004.tb06200.x]

Abbreviations

DOB: date of birth
DP: differential privacy
HIPAA: Health Insurance Portability and Accountability Act
MCBPAARP: Malignant Cancer Big-data Processing Analysis and Application Research Project
PAD: provincial-level administrative division

Edited by G Eysenbach; submitted 15.12.18; peer-reviewed by J Bian, YR Park, A Alaqra; comments to author 17.04.19; revised version received 07.06.19; accepted 26.09.19; published 05.02.20

Please cite as:
Gong M, Wang S, Wang L, Liu C, Wang J, Guo Q, Zheng H, Xie K, Wang C, Hui Z
Evaluation of Privacy Risks of Patients' Data in China: Case Study
JMIR Med Inform 2020;8(2):e13046
URL: https://medinform.jmir.org/2020/2/e13046
doi: 10.2196/13046
PMID: