Research on Privacy-Preserving Methods of Electronic Medical Records

Qingfei Wang\(^1\,^2\), Gen Zhu\(^2\), Changbo Wang\(^3\) and Hongping Cheng\(^1\,^2\,^*\)

\(^1\)Center of Health Administration and Development Studies, Hubei University of Medicine, Shiyan 442000, China
\(^2\)School of Public Health and Management, Hubei University of Medicine, Shiyan 442000, China

\(^*\)Corresponding author e-mail: 936364147@qq.com

Abstract. With the development of Internet technology, electronic medical records (EMRs) are gradually replacing the traditional medical records because of their advantage of portability and ease of storage. At the same time, due to the large number of patient data in EMRs, protecting patients' privacy is an increasing concern that should not be overlooked or understated. This paper studies the privacy preservation mechanism of electronic medical records through two aspects: In the hospital information system, it is analyzed from the policy and regulation, management mechanism and technology level; In the aspect of EMR publishing, this paper mainly introduces K-Anonymity mechanism and differential privacy, which are commonly used privacy protection methods for data publishing.

1. Introduction

With the rapid development of information technology and the advancement of medical and health systems, electronic medical records (EMRs), as the main carrier of medical and health information, have been widely used in hospitals because of their large storage capacity, resource saving, convenient query, and improved efficiency of diagnosis and treatment. Electronic medical records have become a trend in place of traditional paper medical records.

At the same time, the rapid development of EMRs has also brought about the problem of easy disclosure of patient privacy. There are two main ways to disclose medical privacy through EMRs[1-3]: leaking information from hospital internal information system, patient information can be easily leaked because of the easy sharing of EMRs. In some places, there is even a phenomenon in which medical staff sells patient privacy data. In addition, in the process of publishing EMRs, it is easy to cause privacy leakage. For example, in order to collect statistics on patient data, hospitals publish
patient information to medical analysis institutions. If the privacy information is not protected, the patient information will be completely exposed, including the patient's sensitive disease information. At present, the issue of privacy leakage of EMRs has been paid more and more attention. Based on the research of personal privacy protection mechanism, this paper discusses the research status of EMR privacy protection mechanism from two aspects: hospital internal information system and EMRs publishing.

2. Privacy protection mechanism of hospital information system

Privacy protection of hospital medical information mainly includes laws and regulations, management mechanisms and technical means.

2.1 Laws and regulations

In 1996, the United States Government promulgated the Health Insurance Portability and Responsibility Act to strengthen the privacy protection and information security supervision of EMRs. This Act strictly defines the obligations, rights and legal responsibilities of stakeholders in electronic medical records. In 2003, the US government issued the Privacy Standard for Personally Identifiable Health Information, which further strengthens the protection of medical information privacy. The EU also pays great attention to the security and privacy issues in the process of medical information exchange, and has established a digital medical system covering the whole EU. The United Kingdom's Data Protection Act of 1984 stipulates that personal information must be obtained with personal consent. Individuals or institutions holding personal information must have legitimacy and take security measures when using personal information. In China, the use of EMRs was promoted around 2000, but the normative and legal effects of their use were not uniformly regulated. Only a few legal provisions were involved, such as the Practicing physician law, the Nurse management measures, and the medical ethics norms and implementation measures, etc. By 2010, policy documents such as "Basic specifications for electronic medical records (Trial)", "Basic framework and data standards for electronic medical records (Trial)", "Basic specifications for medical record writing" have been issued. However, for specific problems such as EMR authority classification, archive management, medical accident liability determination, no specific operational solutions have been proposed.

2.2 Management mechanism

In terms of management mechanism, hospitals often use access control to protect data privacy. Information resources are classified according to the rights, and users are assigned corresponding rights to access the data, so that data can be used within the legal scope. Access control of EMRs can make the contents of medical records not be accessed by unauthorized users. Reasonable access control of EMRs should be able to be authorized according to the classification of patients and contents of medical records. The representative access control technology is Role-based Access Control (RBAC) proposed by the National Institute of Standards and Technology (NIST). Because the EMR has many functions and can be accessed by many people, such as doctors, nurses, emergency room technicians, and logisticians, it is better to adopt a role-based access control mechanism for medical records to realize role authorization. Different users have different privileges and levels. Role-based access control can reduce the complexity of authorization, reduce management overhead and ensure system security.

2.3 Technical means

At the technical level, the privacy protection of EMRs mainly includes electronic signature technology and data encryption technology.
Since the security of EMRs needs to achieve confidentiality, integrity and non-repudiation, a security mechanism is needed to ensure that patient privacy is not leaked, and electronic signature is the basic technology to achieve this requirement. Electronic signature is a digital signature technology based on Public Key Infrastructure (PKI). The use of electronic signature can effectively ensure the security and integrity of information, as well as the non-repudiation of signature. When using the medical record signature, the full name should be signed, not just the surname; if the name is signed by the intern, the attending physician should re-sign. The use of EMRs can prevent the information in medical records from being tampered with, destroyed and leaked, and make the electronic medical records legitimate. Another way to protect privacy of electronic medical records is data encryption. Especially for some sensitive disease information of patients, the way of information encryption can effectively protect the privacy of patients. Commonly used data encryption technology is symmetric encryption and asymmetric encryption technology. Symmetric encryption is characterized by the use of the same key for encryption and decryption, which is simple and fast, and has a shorter key, such as DES encryption technology; asymmetric encryption requires a pair of keys: public key and private key, one for encryption, and one for decryption. Asymmetric encryption is more secure and difficult to decipher than symmetric encryption technology, such as RSA encryption technology. Data encryption technology for EMRs can protect the privacy of patients with high intensity, but the disadvantage is that the encryption and decryption process has large computational overhead and is not real-time.

3. Privacy protection in EMR publishing

The publication of EMRs has a beneficial effect on medical research, such as the statistical analysis of EMRs by scientific research institutions, which can get the relationship between age and disease, or predict epidemic diseases. Sometimes the hospital also needs to collect patient information and publish it to the medical data center. The medical center conducts statistical analysis of the information, such as statistics on the number of male diabetic patients, or some complex cluster analysis. If the patient's privacy is not protected during the publication of medical records, it will also cause serious privacy disclosure.

In the process of publishing electronic medical records, privacy protection is mainly carried out through technical means. The main methods used are data scrambling, anonymity, generalization, blocking, etc[4]. Among them, data anonymity is a hot research topic in recent years. Typical anonymity protection methods are k-anonymity[5], l-diversity[6] models, and so on.

In the original data of electronic medical records, there are some information that can uniquely identify individual patients, such as telephone number, identity card number, etc., called identifiers; some information can be combined to determine an individual, such as zip code, birthday, gender, which are called quasi-identifiers. There are also some sensitive information, such as patient's disease information. K-anonymity is to hide the identifier information during the publishing process, and then use the generalized quasi-identifier value to make the number of the same group (called the equivalent group) not less than k, so the probability of finding a specific patient record by quasi-identifiers is 1/k, which can effectively protect privacy. The process of protection using the k-anonymity mechanism is shown in the following table:
Table 1: Patient basic information

| ID | Age | Zip code | Disease |
|----|-----|----------|---------|
| 1  | 27  | 225412   | Indigestion |
| 2  | 15  | 225026   | cancer    |
| 3  | 16  | 225132   | cancer    |
| 4  | 24  | 225432   | flu       |
| 5  | 17  | 225322   | cancer    |
| 6  | 18  | 225244   | flu       |

Table 2: Anonymous table satisfying 2-anonymity

| ID | Age | Zip code | Disease |
|----|-----|----------|---------|
| 1  | 2*  | 2254**   | Indigestion |
| 4  | 2*  | 2254**   | flu       |
| 2  | <20 | 225***   | cancer    |
| 3  | <20 | 225***   | cancer    |
| 5  | <20 | 225***   | cancer    |
| 6  | <20 | 225***   | flu       |

Table 1 shows six patient records with hidden identifiers in electronic medical records. By generalizing the quasi-identifier (age, zip code), the 2-anonymity data table 2 is obtained. Publishing Table 2 will not reveal patient's privacy. Because Table 2 satisfies 2-anonymity, i.e. the number of records with the same quasi-identifier in the equivalent group is not less than 2, which means that when linking the anonymous table with the external table, it matches no less than 2 similar records, but can not match the exact individual. L-diversity model is based on k-anonymity, requiring at least L different sensitive attribute values in the same equivalent group. L-diversity is more secure than k-anonymity and can prevent homogeneous attacks, but it can not prevent background knowledge attacks. In recent years, another popular privacy protection algorithm is Differential Privacy, which is a rigorous and provable privacy protection framework that prevents attackers from having arbitrary background knowledge.

Differential privacy aims to maximize the accuracy of data queries while minimizing the opportunity to identify the record when querying from statistical databases. Here is the definition of differential privacy: Let \( \mathcal{A}: D^n \rightarrow Y \) be a randomized algorithm, and \( D_1, D_2 \in D^n \) be two similar datasets, the difference between them is that at most one record is different (called neighboring datasets).

**Definition 1.** Let \( \epsilon > 0 \), define \( \mathcal{A} \) to be \( \epsilon \)-differentially private if for all neighboring datasets \( D_1, D_2 \) and for all (measurable) subsets \( Y \subset Y_k \), we have

\[
\Pr(\mathcal{A}(D_1) \in Y) \leq \exp(\epsilon) \times \Pr(\mathcal{A}(D_2) \in Y)
\]
In general, differential privacy is achieved through Laplace mechanism or Exponential mechanism. First, let \( D^n \rightarrow \mathbb{R}^k \), and \( \| \cdot \|_1 \) be the usual \( L_1 \) norm. Define \( GS(f) \), the global sensitivity of \( f \), for all neighboring datasets \( D_1, D_2 \) as

\[
GS(f) = \| f(D_1) - f(D_2) \|_1
\]  

(2)

Define randomized algorithm \( \mathcal{A} \) as

\[
\mathcal{A}(D) = f(D) + Lap\left( \frac{GS(f)}{\varepsilon} \right)^k
\]  

(3)

Laplace distribution has mean 0, and has density \( p(x; b) = \frac{1}{2b} \exp\left( -\frac{|x|}{b} \right) \).

And \( Lap(b) = (l_1, \ldots, l_k) \), then \( \mathcal{A} \) satisfies \( \varepsilon \)-differential privacy.

Using differential privacy can effectively protect patient privacy. For example, Table 3 is a medical data set in which each record indicates whether a person has cancer (1 means yes, 0 means no). The data set provides users with query services (such as counting queries). Assume that the user can use the query function \( f(i) = count(i) \) to obtain the number of records in the previous i-line that satisfy the "diagnostic result=1". If an attacker knows that Alice is on line 5 of the dataset, then \( count(5) - count(4) \) can be used to infer whether Alice has cancer.

### Table 3. An Example Table of Medical Data Set.

| Name  | Diagnostic result |
|-------|-------------------|
| Tom   | 1                 |
| Jerry | 0                 |
| Lily  | 1                 |
| Henry | 0                 |
| Alice | 1                 |

However, if \( f \) is a function that implements differential privacy protection, such as \( f(i) = count(i) + \text{noise} \), where noise obeys Laplace distribution. Assuming that the possible output of \( f(5) \) comes from the set \( \{3,4,2,1\} \), then \( f(4) \) will also output any value in \( \{3,4,2,1\} \) with almost the same probability, so the attacker cannot get the result through \( f(5) - f(4) \).

Differential privacy solves two drawbacks of traditional privacy protection methods. Firstly, it does not need to consider any background knowledge of the attacker. Secondly, it is based on a solid mathematical foundation. It defines privacy protection strictly and provides a quantitative evaluation method, which makes the level of privacy protection comparable. In the future, we will further study how to apply differential privacy to the privacy protection of electronic medical records.

### 4. Summary

This paper discusses the methods of EMR privacy protection from two aspects. In the hospital internal information system, it analyzes from the policy, management, and technical aspects. In the aspect of electronic medical records publishing, it mainly introduces K-anonymity mechanism, which is commonly used to protect the privacy of data publishing. In short, the privacy protection of electronic medical records needs to be comprehensively considered from various aspects, such as
improving laws and regulations, improving hospital information infrastructure construction, raising awareness of privacy protection, and proposing more optimized privacy protection algorithms, which are the result of joint efforts in many aspects.

Acknowledgment

Fund Project: This paper is supported by Key Research Center for Humanities and Social Sciences in Hubei Province（Hubei University of Medicine）

References

[1] Kahn S, Sheshadri V. Medical record privacy and security in a digital environment[J]. IT Professional, 2008,10(2):46-52.
[2] Kim D M, Hong B W. Personal privacy protection of Electronic Medical Record in health information system [J]. Journal of Korean Society of Medical Informatics, 2007,13(3):249-257.
[3] Dean B B, Lam J, Natoli J L, et al. Review: Use of Electronic Medical Records for health outcomes research a literature review[J]. Medical Care Research and Review, 2009,66(6):611-638.
[4] Fung B, Wang K, Chen R, et al. Privacy-preserving data publishing: A survey of recent developments[J]. ACM Computing Surveys(CSUR), 2010,42(4):14.
[5] Sweeney L. K-anonymity: A model for protecting privacy[J]. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 2002,10(05):557-570.
[6] Machanavajjhala A, Kifer D, Gehrke J, et al. L-Diversity: Privacy beyond K-Anonymity[J]. ACM Transactions on Knowledge Discovery from Data(TKDD),2007,1(1):3.
[7] Tran D, Sokolova M. Applying multi-label and multi-class classification to enhance K-anonymity in sequential releases[J]. Progress in Artificial Intelligence. Berlin: Springer,2016,5(4):277-288.
[8] Lau E C, Mowat F S, Kelsh M A, et al. Use of electronic medical records (EMRs) for oncology outcomes research: assessing the comparability of EMR information to patient registry and health claims data[J]. Clinical Epidemiology, 2011, 3(3):259-272.
[9] Loukides G, Gkoulalasdivanis A, Malin B. Anonymization of electronic medical records for validating genome-wide association studies[J]. Proceedings of the National Academy of Sciences, 2010, 107(17):7898-7903.
[10] Wang Z H, XU J, Wang W, et al. Clustering-based approach for data anonymization[J]. Journal of Software, 2010, 21(4): 680-693.
[11] Diesburg S M , Wang A I A . A survey of confidential data storage and deletion methods[J]. ACM Computing Surveys, 2010, 43(1):1-37.
[12] Dwork C . A firm foundation for private data analysis[J].Communications of the ACM, 2011,54(1):86-95.
[13] Soria-Comas J, Domingo-Ferrer J, Sánchez D, et al. t-Closeness through Microaggregation: Strict Privacy with Enhanced Utility Preservation[J]. IEEE Transactions on Knowledge and Data Engineering, 2015, 27(11):3098-3110.
[14] Raymond Wong, Jiuyong Li, Ada Fu, et al. ( α, k )-anonymous data publishing[J]. Journal of Intelligent Information Systems, 2009, 33(2):209-234.