Chapter 53
Framework Design and Case Study for Privacy-Preserving Medical Data Publishing

Yu Niu
Tsinghua University, China

Ji-Jiang Yang
Tsinghua University, China

Qing Wang
Tsinghua University, China

ABSTRACT

With the pervasive using of Electronic Medical Records (EMR) and telemedicine technologies, more and more digital healthcare data are accumulated from multiple sources. As healthcare data is valuable for both commercial and scientific research, the demand of sharing healthcare data has been growing rapidly. Nevertheless, health care data normally contains a large amount of personal information, and sharing them directly would bring huge threaten to the patient privacy. This paper proposes a privacy preserving framework for medical data sharing with the view of practical application. The framework focuses on three key issues of privacy protection during the data sharing, which are privacy definition/detection, privacy policy management, and privacy preserving data publishing. A case study for Chinese Electronic Medical Record (ERM) publishing with privacy preserving is implemented based on the proposed framework. Specific Chinese free text EMR segmentation, Protected Health Information (PHI) extraction, and K-anonymity PHI anonymous algorithms are proposed in each component. The real-life data from hospitals are used to evaluate the performance of the proposed framework and system.

1. INTRODUCTION

As more and more health care information systems adopted by medical institutions, health care data have been accumulated rapidly in the past decades (Kimberly, 2012). The efficiency of medical institution could be improved and the cost could be reduced by using and sharing electronic healthcare data (Open clinical 2005; Samuel 2003). A review by Dr. Piwowar indicates that the healthcare data

DOI: 10.4018/978-1-4666-8111-8.ch053

sharing is crucial to academic health centres for research too (Healther, 2008).

However, the widely usage and sharing of health care data have bring in many concerns. With detailed person-specific data contained in healthcare data, sensitive information about individuals may be easily revealed by analyzing the shared data. Research shows that patients could be easily identified by using identifiers or specific combined information (such as age, address, gender) from each other in a certain health care dataset. An example in Samarati (2001) shows that linking medication records with voter lists can uniquely identify a person's name and his/her medical information. As a result of these patient privacy leakage concerns, privacy protection laws are passed in many counties. In 1996, the Health Insurance Portability and Accountability Act (HIPAA) title II was enacted in US. One of the purposes of this act is to increase the protection of patients' medical records against unauthorized usage and disclosure. In 1997, Council of Europe announced the Recommendation R (97) 5 on the Protection of Medical Data to enhance the protection of personal health care data. Similar regulations have been enacted in many other countries.

Many approaches have been proposed in the past years to eliminate the privacy concerns from patients and help medical institutions or participants to comply with those privacy protection regulations. Most of them are focusing on a small scope of the problem with single theory. Based on the analysis of EMR publishing procedure, the privacy protection execution can have three stages. The first one is the definition of privacy. The second one is how to express and manage the complex requirements by privacy laws and regulations in the information system. The third one is how to publish the health data based on the given privacy definition and policy.

To cope with the above three issues, a practical privacy preserving data sharing framework is proposed. The framework consists of three layers. The private data definition/detection component and privacy policy definition component are in the bottom layer as the base of the whole framework. The policy management in the middle layer enforces the resolved privacy policy on the medical data. The data access control and publishing component in the top layer provide direct service for data sharing.

A case study for privacy-preserving medical data publishing is introduced as the implementation of the framework. In the case, EMRs are processed to eliminate the private and sensitive parts before being published. During the implementation, free text Chinese EMR word segmentation, Protected Health Information (PHI) extraction, and narrative text K-anonymity algorithm are contributed to realize the privacy detection and EMR anonymous publishing functions of the framework. Real-life data from hospitals are used to evaluate the proposed framework. To the best of our knowledge, this is one of the few research works about privacy preserving framework on the real medical data sharing.

The rest of this paper is arranged as follow. The related literature works are reviewed and discussed in Section 2. Section 3 proposes a medical data privacy preserving framework with the fundamental design rules and guidelines for each component. In Section 4, the detailed privacy detection and record anonymity algorithms are explained and implemented in a case study of free text Chinese EMR publishing system. Experiment results are analyzed and evaluated in Section 5. With the implementing and experiment results, a conclusion and future work plan are given in Section 6.

2. RELATED WORKS

2.1. Data Sharing

Data sharing could happen in scale of small amount or massive two ways. The corresponding privacy protection challenges and solutions are quite different. And for data sharing, the associated privacy and policy management is also necessary (Figure 1).