Chapter 7

Privacy Preserving Integration of Health Care Data

Xiaoyun He
Rutgers University, USA

Jaideep Vaidya
Rutgers University, USA

Basit Shafiq
Rutgers University, USA

Nabil Adam
Rutgers University, USA

Tom White
NY Office of Mental Health & Columbia University, USA

ABSTRACT

For health care related research studies the medical records of patients may need to be retrieved from multiple sites with different regulations on the disclosure of health information. Given the sensitive nature of health care information, privacy is a major concern when patients’ health care data is used for research purposes. In this paper, the authors propose approaches for integration and querying of health care data from multiple sources in a secure and privacy preserving manner. In particular, the first approach ensures secure data integration based on unique identifiers, and the second one considers data integration based on quasi identifiers, for which a rule-based framework is proposed for cross-linking data records, including secure character matching.

INTRODUCTION

Health care research studies often involve analysis of huge amount of data collected from various sources including health care providers, pharmacies, insurance companies, government agencies, and research institutions. Given the sensitive nature of health information and the social and legal implications for its disclosure, privacy is a major concern for information sharing in the healthcare domain (Rindfleisch, 1997; Kelman et al., 2002; HIPAA, 2000).

Protecting the privacy of individually identifiable health information is more important when such information is used for clinical or health services related research. The Health Insurance Portability and Accountability Act (HIPAA) privacy rule strictly prohibits sharing of individu-
ally identifiable health information with clinical researchers who are not covered entities. The covered entities, as defined in this privacy rule, include health plans, healthcare clearing houses, and healthcare providers that transmit health information electronically in connection with certain defined HIPAA transactions, such as claims or eligibility inquiries (HIPAA, 2000). For research purposes, only de-identified or anonymized health information can be used.

Several National initiatives are addressing the privacy and security concerns raised by HIPAA. The Health Information Security and Privacy Consortium (HISPC) is documenting and reconciling differences in state, local, and federal privacy and security laws with a goal of enabling privacy-protected sharing of data across state lines, and so that those data can be incorporated into a National Health Information Network (NHIN). HISPC must reconcile differences between HIPAA, and numerous state-specific laws regulating sharing of mental health, substance abuse, HIV, and cancer information. In some states, conflicting regulations prevent the sharing of clinical data about mental health and substance abuse within single hospitals, even though many patients have both mental health and substance abuse diagnoses. Successful completion of the HISPC efforts is essential to establish the trust needed to motivate consumers to share their data through the NHIN. In this paper, we present a technique for integrating data across organizations that is of direct relevance to these efforts.

As discussed above, for health care related research studies the medical records of patients may need to be retrieved from multiple sites with different regulations on the disclosure of health information. In absence of the identity information, correlation and integration of such records on a per patient basis in a privacy preserving manner is an important research issue. As an example, consider the following queries related to a research study for determining defective anti-depressant drugs:

**Query 1:** What percentage of HIV infected patients taking any prescribed anti-depressant medication are diagnosed with acute psychiatric disorder?

**Query 2:** For each HIV infected patient diagnosed with acute psychiatric disorder find all the prescribed drugs the patient took after being diagnosed with HIV.

The above queries require integrating data from multiple sources including the state health department managing HIV test records, pharmacy databases storing patient records related to prescription drugs, and mental health clinics treating patients with psychiatric disorder. For preserving the privacy of individual records, we need to ensure that the query result do not reveal any individually identifiable information to the querying party. Additionally, during the process of integrating data from multiple sources, none of the sources should be able to learn/infer any information about any of the patients beyond what these sources already know. For instance, the pharmacist should not be able to learn which patients have been tested positive for HIV or which patients are receiving treatment for severe psychiatric disorder.

One of the large challenges in merging data is the lack of a common identifier across data systems. There are numerous commercial applications for creating enterprise master person indexes from distributed databases. Some of these are tolerant of missing, mistyped, or conflicting data. The approach commonly used by Regional Health Information Organizations (RHIOs) is that of RxHub – to only cross-link patients who have exact matches of five elements: first name, last name, birthday, gender, and zip code. RxHub is a third party data aggregator which has contracts with most major pharmacy benefits managers. Hospitals may use RxHub to request all historical pharmacy data from patients by sending those five demographic parameters to RxHub. RxHub