Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 98 (2016) 324 – 331

Procedia Computer Science

The 6th International Conference on Current and Future Trends of Information and Communication Technologies in Healthcare (ICTH 2016)

Querying medical datasets while preserving privacy

Nafees Qamar[a], Yilong Yang[b], Andras Nadas[c], Zhiming Liu[d,*]

[a]*College of Computer Science and Engineering, University of Hafr Al Batin*
[b]*Department of Computer and Information Science, Faculty of Science and Technology, University of Macau*
[c]*Institute for Software Integrated Systems, Vanderbilt University*
[d]*Southwest University, China*

Abstract

This paper addresses the challenge of identifying clinically-relevant patterns in medical datasets without endangering patient privacy. To this end, we treat medical datasets as *black box* for both internal and external users of the data enabling a remote query mechanism to construct and execute database queries. The novelty of the approach lies in avoiding the complex data *de-identification* process which is often used to preserve patient privacy. The implemented toolkit combines software engineering technologies such as Java EE and RESTful web services, to allow exchanging medical data in an unidentifiable XML format along with restricting users to the *need-to-know* privacy principle. Consequently, the technique inhibits retrospective processing of data, such as attacks by an adversary on a medical dataset using advanced computational methods to reveal Protected Health Information (PHI). The approach is validated on an endoscopic reporting application based on openEHR and MST standards. The proposed approach is largely motivated by the issues related to querying datasets by clinical researchers, governmental or non-governmental organizations in monitoring health care services to improve quality of care.

Peer-review under responsibility of the Program Chairs

Keywords: Data disclosure; Data de-identification; Web Services; Data Privacy; Automated Software Engineering

1. Introduction

Patients' Electronic Health Records (EHRs) are stored, processed, and transmitted across several healthcare platforms and among clinical researchers for on-line diagnostic services and other clinical research. This data dissemination serves as a basis for prevention and diagnosis of a disease and other secondary purposes such as health system planning, public health surveillance, and generation of anonymized data for testing. However, exchanging data across organizations is a non-trivial task because of the embodied potential for privacy intrusion. Medical organizations tend to have confidential agreements with patients, which strictly forbid them to disclose any identifiable information of the patients. Health Insurance Portability and Accountability Act (HIPAA) explicitly states the confidentiality protection on health information that any sharable EHRs system must legally comply with. To abide by these strict regulations,

*Corresponding author. Tel.: +86-(0)23-68367358 ; fax: +86-(0)23-68367358.
E-mail address: zhimingliu88@swu.edu.cn

Peer-review under responsibility of the Program Chairs
doi:10.1016/j.procs.2016.09.049

data custodians generally use de-identification techniques[21 11 20] so that any identifiable information on patient's EHR can be suppressed or generalized.

However, in reality, research[19] indicates that 87% of the population of U.S. can be distinguished by sex, date of birth and zip code. We can define quasi-identifiers as the background information about one or more people in the dataset. If an adversary has knowledge of these quasi-identifiers, it can possibly recognize an individual and take advantage of his clinical data. On the other hand, we can find out most of these quasi-identifiers have statistical meanings in clinical research. There exists a paradox between reducing the likelihood of disclosure risk and retaining the data quality. For instance, if information related to patients' residence was excluded from the EHR, it would disable related clinical partners to catch the spread of a disease. Thus, strictly filtered data may lead to failure in operations. Conversely, releasing data including patients' entire information including residence, sex and date of birth would bring a higher disclosure risk.

In this paper we address the emerging problem of de-identification techniques, namely, the problem of offering de-identified dataset for a secondary purpose that makes it possible for a prospective user to perform retrospective processing of medical data endangering patient privacy. Our approach differs from the traditional techniques in the sense that it employs software engineering principles to isolate and develop key requirements of data custodians and requesters. We apply Service-Oriented Architecture (SOA) that provides an effective solution for connecting business functions across the web—both between and within enterprises[8] Our proposed toolset integrally relies on web services. The results are retrieved in an XML data format that excludes all personal information of patients. The basic model used here follows the principles of RESTful web services by combining three elements: a *URLs repository* for identifying resources uniquely corresponding to clinical data queries, *service consumers* requesting data, and *service producers* as custodians of clinical data. The implemented toolkit uses Java EE that offers an easy way to develop applications using EJBs. Needless to mention that Java EE is widespread and is largely used by community. Our proof-of-concept implementation uses GastrOS, an openEHR[7] database[1] describing an endoscopic application. Our toolset enables answering queries such as: *Find the number of patients who are still susceptible to developing a Hepatitis B infection even after full compliance to the Hepatitis B vaccination schedule–i.e. the baseline and second detection dates for the HBsAg and Anti-HBs tests both show negative results.*

The set of clinical data queries described in the paper have been crafted with the help of clinical researchers at Vanderbilt University. Our approach mainly contributes to the development of privacy preserving techniques on patient data by treating datasets as *blackboxes*. In this way, disclosure risks associated with patient data are minimized.

The paper proceeds as follows: Section 2 describes the related work; Section 3 states an application example; Section 3.1 presents the technical details of our approach; Section 4 overviews the clinical data queries corresponding to the GastrOS dataset; Section 5 discusses the authentication and authorization mechanism connecting users to clinical datasets; Section 6 summarizes the work and details some future research directions.

2. Related Work

In contrast to some of the existing techniques[12 13] our approach relies on advanced software engineering principles and technologies for analyzing clinical datasets. For example, caGrid 1.0[12] (now caGrid 2.0), released in 2006, is an approach that discusses a complex technical infrastructure for biomedical research through an interconnected network. Similar work in[2] discusses a combined interpretation of biological data from various sources. This work, however, considers the problem of continuous updates of both the structure and content of a database and proposes the novel database SYSTOMONAS for SYSTems biology of pseudOMONAS. Interestingly, this technique combines a data warehouse concept

De-identification techniques for medical data have been studied and developed by statisticians dealing with integrity and confidentiality issues of statistical data. The major techniques used for data de-identification are (i) CAT (Cornell Anonymization Kit)[21], (ii) μ-Argus[11], and (iii) sdcMicro[20]. CAT anonymizes data using generalization, which is proposed[3] as a method that specifically replaces values of quasi-identifiers into value ranges. μ-Argus is an acronym for Anti-Re-identification General Utility System and is based on a view of safe and unsafe microdata that

[1] `http://gastros.codeplex.com`